REFERENCE ONLY

UNIVERSITY OF LONDON THESIS

Degree PhD Year 2006 Name of Author Woods, C.A

The regulation of PI3-kinase p110δ gene expression

A dissertation submitted to the University of London in candidature for the degree of Doctor of Philosophy

By

Elizabeth Woods

Department of Biochemistry and Molecular Biology
&
Ludwig Institute for Cancer Research

University College London

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Phosphoinositide 3-kinases (PI3Ks) are a family of signal transduction enzymes which generate lipid second messengers (1) and control a wide variety of cellular processes such as growth, proliferation, survival, differentiation, intracellular traffic, cytoskeletal organisation and cell migration (1-7)

Mammals have 8 isoforms of PI3K, most of which have poorly defined individual roles in cells and in the organism. Deregulated PI3K signalling has been implicated in cancer, inflammation and diabetes, and PI3Ks are being pursued as new therapeutic targets by the pharmaceutical industry. Therapeutic intervention with PI3K will almost certainly have to be targeted at individual isoforms, given the risk that global alteration of PI3K signalling will be deleterious to the organism.

This work focuses on the p110δ isoform of PI3K. p110δ belongs to the class IA subset of PI3Ks which signal downstream of tyrosine kinase receptors and Ras. These PI3Ks are heterodimers that consist of a p110 catalytic subunit and a regulatory subunit. Mammals have genes for 3 catalytic subunits, called p110α, p110β and p110δ. Whereas p110α and p110β show a broad tissue distribution, p110δ is more restricted, with the highest levels in leukocytes (8) There is mounting evidence for a non-redundant function of these p110 isoforms (7,9). Studies using mice have implicated p110δ in immune signalling, making it an interesting new target for anti-inflammatory therapies. More recent work (7) also indicates a role for p110δ in cancer.

This thesis studies the mechanism of tissue specific distribution of p110δ at the mRNA and protein level. Using Real-Time reverse transcriptase PCR we have demonstrated that regulation of p110δ expression occurs at the level of transcription. Using two different approaches, 5' RACE (Rapid Amplification of cDNA Ends) and reverse transcriptase PCR, we have also discovered different transcription start sites in the p110δ mRNA, leading to three discrete mRNAs which contain distinct 5' untranslated exons. In a further stage of the work, we

have found evidence that differential usage of these 5' untranslated exons correlates with the distinct tissue distribution of the p110δ protein.

Our findings imply that there are at least 3 distinct promoters for the p110δ gene. Efforts to delineate the promoter elements using reporter assays and deletion analysis have given ambiguous results, and have not lead to the desired insight into the regulation of the p110δ promoters and their regulation in different tissues.

Acknowledgements

There are many people to whom I am indebted for their help in the preparation of this thesis and without whom I would not have been able to complete it.

Chief amongst these is Dr Bart Vanhaesebroeck the academic and intellectual scaffolding for the project.

Dr Klaus Okkenhaug instigated the search for the mode of regulation of the PI3K p110δ protein.

Dr Klaartje Kok provided valuable assistance for experiments in Chapters 3 and 4.

The project was funded by the BBSRC and The Ludwig Institute for Cancer Research

I would like to thank all members of the Cancer Cell Signalling Group for their friendship over the last four years. Finally and by no means least I am grateful to my wonderful husband and family for their moral support.

Table of Contents

List of Tables

List of Figures

Abbreviations

ACK	ammonium, chloride, potassium
Amp	ampicillin
APS	ammonium persulphate
ATP	adenosine triphosphate
bp	base pair
cDNA	complimentary DNA
CIAP	calf intestine alkaline phosphatase
cpm	counts per minute
CPSF	cleavage and polyadenylation specificity factor
CstF	cleavage stimulatory factor
CTP	cytidine triphosphate
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
DEPC	diethylpyrocarbonate
dGTP	deoxyguanosine triphosphate
dH_2O	distilled water
ddH_2O	double distilled water
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide-triphosphate
DMEM	Dulbeccos modified Eagles medium
DMSO	dimethyl sulfoxide
dTTP	deoxythymidine triphosphate
DTT	dithiothreitol
ECL	enhanced chemiluminescence
EDTA	ethylenediaminetetraacetic acid
ESE	exonic splicing enhancer
ESS	exonic splicing silencer
EST	expressed sequence tag
FCS	foetal calf serum
GAP	GTPase activating protein
GEF	guanosine nucleotide exchange factor
GTP	guanosine triphosphate
HCl	hydrogen chloride
HRP	horse radish peroxidase
HSS	hypersensitive sites
IAA	isoamyl-alcohol
IP	immunoprecipitation
IPTG	isopropyl-1-thio-β-D-galactoside
kb	kilobase pairs
LB	Luria-Bertani (medium)
LCR	locus control region
MOPS	3-(N-morpholino)propanesulphonic acid
OD	optical density
PABP	poly-A binding protein
PAGE	polyacrylamide gel electrophoresis
PAP	poly-A polymerase

PCR	plolymerase chain reaction
PBS	phosphate buffered saline
PDGFR	platelet derived growth factor receptor
PH	pleckstrin homology
PI	phosphoinositides
PI3K	phosphoinositide 3 kinase
PMSF	phenylmethylsulphonyl fluoride
RACE	rapid amplification of cDNA ends
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
RPMI-1640	Roswell Park Memorial Institute
RRM	RNA recognition motif
RT	reverse transcriptase
RT-PCR	reverse transcriptase polymerase chain reaction
SDS	sodium dodecyl sulphate
snRNP	small nuclear ribonucleoprotein
TAE	tris-acetate-EDTA
TAF	TBP-associated factor
TAP	tobacco acid pyrophosphatase
TBE	tris/borate buffer
TBP	TATA binding protein
TBS	tris-buffered saline
TBST	tris-buffered saline tween
TE	tris/EDTA buffer
TEMED	tetramethylethylenediamine
TF	transcription factor
TFBS	transcription factor binding site
TLCK	Na-p-tosyl-L-lysine chloromethyl ketone
UTP	uridine triphosphate
UTR	untranslated region
UV	ultra-violet
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside

1 Introduction

1.1 General aspects of gene regulation

In eukaryotic cells, the ability to express biologically active proteins is regulated at several levels: Regulation of chromatin structure, transcriptional initiation, transcript processing and modification, RNA transport, protein post-translational modification, protein transport and the control of protein stability. These processes will be described briefly below:

1.1.1 Chromatin Structure

The physical structure of the DNA can affect the ability of transcriptional regulatory proteins, transcription factors, and RNA polymerases to access specific genes and to activate their transcription. The accessibility of the chromatin to RNA polymerases and transcription factors is most affected by the presence of histones and CpG methylation (10,11).

Due to the large amount of DNA in each eukaryotic cell, compaction of the large DNA molecules is achieved by interaction of the DNA with proteins to form the protein–DNA complex called chromatin (10). Different types of chromatin have different levels of packaging. Transcribed chromatin has an extended structure whereas inactive (not transcribed) chromatin has a more condensed structure (Figure 1-1). Examination of eukaryotic nuclei under a microscope shows that there are two different condensation states of chromatin in the nucleus. Heterochromatin is highly condensed (about 10% of the nucleus) and probably does not contain many active genes. Euchromatin is much less condensed and about 10% of the euchromatin is believed to consist of genes that are either actively transcribed or poised for transcription (11).

There are two broad classes of chromatin-binding proteins.

1. **Histones** - This class of DNA-binding proteins includes four different proteins, H2A, H2B, H3 and H4. All histones are small, very basic

proteins rich in lysine and arginine. The histones are the basic building blocks of chromatin structure.

2. **Non-histone chromosomal proteins** – This is a much more diverse group of DNA-binding proteins which include polymerases, other nuclear enzymes, hormone receptor proteins, and regulatory proteins of many kinds. Approximately 1000 different non-histone chromosomal proteins are known in a typical eukaryotic nucleus (12).

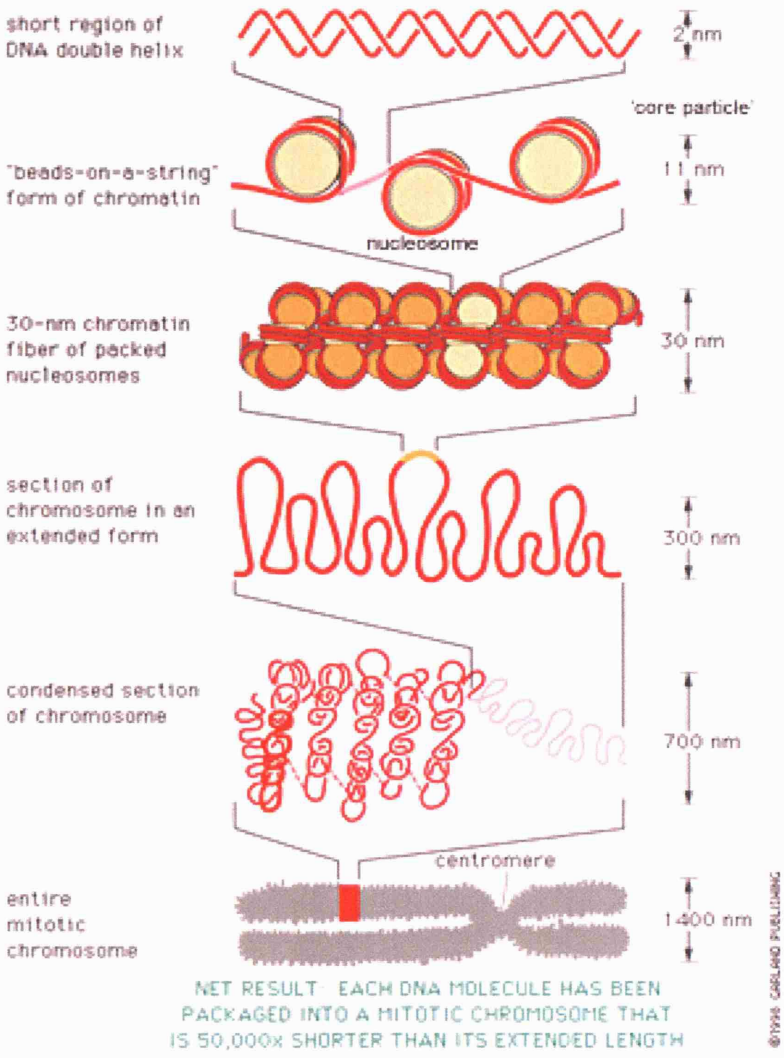

Figure 1-1 Levels of chromatin packing. This schematic drawing shows some of the orders of chromatin packing thought to give rise to the highly condensed mitotic chromosome. The folding of naked DNA into nucleosomes is the best understood level of packing. Due to the large amount of DNA in each eukaryotic cell, compaction of the large DNA molecules is achieved by interaction of the DNA with proteins to form the protein–DNA complex called chromatin. Different types of chromatin have different levels of packaging. Transcribed chromatin has an extended structure whereas inactive (not transcribed) chromatin has a more condensed structure. Figure taken from Alberts *et al.* (11).

The precise role of histones was not understood until about 1974. Then, research in a number of laboratories showed that these proteins combine in a specific way with DNA to form a repeating element of chromatin structure, called the nucleosome (13).

Nucleosomes always contain 146 bp of DNA, wrapped around an octamer of histone molecules. A histone H3-H4 tetramer interacts with two H2A-H2B dimers (14). This complex is known as the "core particle" of the nucleosome (Figure 1-1). (15). The DNA lies on the surface of the histone octamer and makes about 1.75 left-hand super helical turns around it. The structure of the octamer provides a platform upon which the DNA is bound. A single molecule of histone H1 constrains the linker DNA between adjacent nucleosomes (16).

In recent years it has been shown that the nucleosome is an integral component and regulator of transcriptional mechanisms. Activator proteins bound to enhancer DNA elements stimulate the initiation of transcription. The activators must gain access to their recognition sequence before this can take place and their access is limited by packaging of these sequences into nucleosomes (17-20). Remodeling of chromatin, possibly involving removal or repositioning of nucleosomes, accompanies transcriptional activation (21-23). Studies have shown that the remodeling process can be mediated by special multiprotein complexes that function at diverse promoters (23). Once an activator protein has gained access and bound to a site in the chromatin, it may cause further displacement of histones, clearing the way for additional activator or other protein binding. Histone displacement leads to the formation of nucleosome-free regions in chromosome, known as DNase hypersensitive sites (24).

All known chromatin modifying and remodeling complexes target the nucleosome, either through acetylation/deacetylation of histone tails or by disrupting histone-DNA interactions (25,26).

1.1.1.1 ATP-dependent chromatin remodeling

ATP-dependent chromatin remodeling complexes alter the structure of chromatin during gene regulation by introducing a superhelical negative twist into DNA. All these complexes contain ATPase subunits, belonging to the SNF superfamily of proteins, which use the energy of ATP hydrolysis to disrupt or alter the association of histones with DNA. However, the exact means by which they do this remains unclear (25).

The most well studied ATP-dependent transcriptional activators are SWI2/SNF2 and SNF5 belonging to yeast. Early work showed that the chromatin from *snf2* and *snf5* mutants had a different nuclease cleavage pattern from that of wild-type yeast. SNF2 also has homology to known helicases, an activity which is required for their function (23). *snf2* and *snf5* mutations caused defects in transcription of SUC2, a glucose-repressible gene encoding the enzyme invertase (27), demonstrating their role in transcription activation.

DNA microarray analysis of global gene expression in yeast identified the SWI/SNF complex as being involved in transcriptional repression. Inactivation of the remodeling complex altered the expression levels of 6% of all yeast genes, most of them negatively (28). The direct or indirect influence of the SWI/SNF complex on these genes is not known. Other members of the SNF superfamily of proteins include the *Drosophila* Brahma complex, mouse mBRG1 and human hBRM and hBRG1 (25).

Some remodeling factors seem to be targeted to specific regions of the genome (29). This makes the order of recruitment slightly less clear. Does the DNA remodeling factor facilitate access of the DNA binding protein to the DNA or does the DNA binding factor recruit the remodeling factor? There is evidence for both scenarios. The Gal 4 is a transcriptional activator that activates genes involved in galactose catabolism. Gal 4 can bind to DNA packaged into chromatin, it also has been shown to bind SWI/SNF (30). This evidence is consistent with the 'recruitment of the remodeling factor' scenario. Other factors such as CTF/NF-1 (CCAAT transcription factor/nuclear factor 1), a transcription

factor that participates in the transcriptional activation of numerous cellular and viral genes, bind to their recognition sites in chromatin (31) therefore need the chromatin to be disrupted prior to their recruitment.

1.1.1.2 Methylation

CpG sites are regions of the DNA where a cytosine nucleotide is situated next to a guanine nucleotide. CpG stands for cytosine and guanine separated by a phosphate which links the two nucleotides together in DNA. Despite the high probability of a cytosine and guanine lying next to each other in a random nucleotide sequence, there are surprisingly few CpG sites in the eukaryotic genome. In vertebrates 60-90% of all CpGs are methylated which leaves only a small part of the genome methylation free. DNA methyltransferase methylates the cytosine in the CpG site forming 5' methylcytosine. Spontaneous deamination converts 5-methylcytosine into 5' methyluracil. If the mutation has no biological impact, as in most cases, the repair machinery does not correct the error. This then results in the loss of the CpG site in the genome. A high concentration of CpG sites in the DNA constitutes what is called a CpG island. Many of the remaining non-methylated CpGs are found in CpG islands (~ 15% of all CpGs in human DNA). The lack of methylation of these particular CpGs means that the repair machinery recognizes any spontaneous deamination of cytosine to uracil and the CpG site is restored. These islands usually include functional promoters (32). Early experiments demonstrated that artificially methylated DNA can adopt a distinctive chromatin structure upon integration into the genome (33). Once formed into chromatin structures the DNA becomes resistant to nuclease digestion which leads to the loss of DNase-I-hypersensitive sites and a chromatin structure similar to that of the bulk of the methylated genome. In contrast, unmethylated CpG islands, which are associated with promoters of many genes, have a more relaxed, nuclease-sensitive chromatin structure (34).

The most direct mechanism by which DNA methylation could interfere with transcription would be by preventing the basal transcriptional machinery and transcription factors to bind to the promoters. This was thought to be the case for

many years until Bird and his colleagues identified MeCP1 and MeCP2, specific transcriptional repressors that recognise methyl-CpG (35,36). MeCP2 contains a methyl CpG DNA-binding domain for direct chromatin structure alteration as well as a repressor domain for long-range repression (37).

Genetic experiments have demonstrated that proper control of DNA methylation is essential for normal mammalian development. It has been shown to play a role in X chromosome inactivation, senescence and carcinogenesis (38-40). Until recently the molecular mechanism by which DNA methylation represses transcription was unknown. Recent reports have linked transcriptional repression to DNA methylation in connection with histone deacetylation (41). MeCP2 has been found to repress transcription via the recruitment of histone deacetylases to methylated DNA regions (42). This leads to a denser packing of the chromatin and generates and unfavourable environment for transcription.

1.1.1.3 Histone deacetylation

An inhibitor of histone deacetylases (HDAC), sodium butyrate, was shown to increase the expression of a methylated episomal reporter gene (43). It was also reported that the use of both sodium butyrate and Trichostatin A (TSA), a specific inhibitor of HDACs restore transcription to previously methylated and silenced plant ribosomal genes. The same result as produced by 5-aza-2'deoxycytidine (5-aza-dC), a DNA demethylating agent (44). It was these clues that prompted the notion that DNA methylation and histone deacetylation operate along a common pathway to silence transcription. Acetylation of histones involves the transfer of an acetyl group from acetyl coenzyme A to the ε-amino group of a lysine residue at the N-terminal region of the histone. This modification reduces the net positive charge of the core histones, leading to a decrease in their binding affinity for DNA. The ends of the histones are then displaced from the nucleosome, the nucleosome unfolds and provides access for transcription factors (45). Therefore when MeCP2 recruits HDAC to methylated regions of DNA, the histones H3 and H4 are deacetylated generating an unfavourable environment for transcription.

Covalent modifications of the amino-termini of the core histones in nucleosomes have been shown to be one of the key regulatory mechanisms in transcription regulation. Recently, new histone modifications have been uncovered besides acetylation, including phosphorylation, methylation and ubiquitination (20).

1.1.2 Transcription

This is a key level of control of gene expression in eukaryotes. Specific factors that exert control include the strength of promoter elements within the DNA sequences of a given gene, the presence or absence of enhancer sequences (which enhance the activity of RNA polymerase at a given promoter by binding specific transcription factors), and the interaction between multiple activator proteins and inhibitor proteins (46). There is strong evidence that suggests a large number of genes that are controlled via multiple promoters exhibit tissue-specific expression from one or more of the distinct promoters (47-53).

Initiation of transcription is a vital step in gene expression. Without the initiation of transcription, and the subsequent transcription of the gene into mRNA by RNA polymerase, the phenotype controlled by the gene will not be seen.

The control of transcription is a delicate mechanism involving cis-acting sequences and trans-acting factors. Cis-acting sequences usually lie 5' of the transcriptional start site. These sequences are the substrate for trans-acting factors. These factors bind to the cis-acting sequences and prepare the DNA in their vicinity for transcription. Trans-acting factors are proteins and must therefore also be encoded by genes. This interplay between genes and their cis-acting sequences and trans-acting factors is a cascade of genetic events (11,46).

Transcription is the mechanism by which a template strand of DNA is utilized by specific RNA polymerases to generate one of the three different classifications of RNA. These 3 RNA classes are:

Class I. Messenger RNAs (mRNAs): This class of RNA is the template for the genetic code used by the translational machinery to determine the order of amino acids incorporated into an elongating polypeptide in the process of translation.

Class II. Transfer RNAs (tRNAs): This class of small RNAs form covalent attachments to individual amino acids and recognize the encoded sequences of the mRNAs to allow correct insertion of amino acids into the elongating polypeptide chain.

Class III. Ribosomal RNAs (rRNAs): This class of RNA assembles together with numerous ribosomal proteins to form the ribosomes. Ribosomes engage the mRNAs and form a catalytic domain into which the tRNAs enter with their attached amino acids. The proteins of the ribosomes catalyse all of the functions of polypeptide synthesis (46).

All RNA polymerases are dependent upon a DNA template in order to synthesize RNA. One strand of DNA holds the information that codes for various genes; this strand is often called the template strand or antisense strand (containing anticodons). The other, and complementary, strand is called the coding strand or sense strand (containing codons). Since mRNA is made from the template strand, it has the same information as the coding strand. However, in RNA, U is substituted for T (11,46).

1.1.3 Transcription initiation

In eukaryotes, a DNA-dependent multisubunit enzyme, RNA polymerase II, controls transcription initiation. This enzyme cannot act alone as it is unable to recognize promoter sequences directly, and cannot to respond to external signals required for the tissue-specific or temporally specific expression of certain genes. Therefore accurate initiation requires the assembly of general initiation factors (54).

Transcription initiation at eukaryotic protein encoding genes is preceded by the assembly of general initiation factors, TF (transcription factor) IIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, on the promoter along with the RNA polymerase II thus creating a large multi-protein DNA complex, the preinitiation

complex (described in more detail in 1.1.3.2). Environmental and developmental signals are dealt with by yet another group of regulatory factors, transcriptional activator and coactivators (28). Protein-coding gene promoters contain a number of DNA sequences involved in transcriptional regulation; these include core or basal promoter elements, promoter proximal elements and distal enhancer elements (described in more detail in section 1.1.3.3). Transcription factors work together with these three types of DNA target and with each other to maintain tight control over transcription initiation (55).

1.1.3.1 Promoters

The promoter is the term used to describe all the sequences that are important in initiation of transcription of a gene. Some genes have numerous and diverse promoters, including the core/basal promoter, the site at which the initiation complex is assembled, as well as one or more upstream promoter elements such as CCAAT boxes and enhancers. Assembly of the initiation complex on the core promoter can usually occur in the absence of the upstream elements, but only in an inefficient way. This suggests that the proteins that bind to the upstream elements include some activators of transcription, and which therefore 'promote' gene expression (46).

a) TATA box promoters

TATA elements are the most common and best characterized core promoter elements. In general, specialized genes that encode proteins made only in certain types of cells have a TATA box. These elements are found 25 base pairs (bp) upstream of the transcription start sites and have a consensus sequence TATAa/tAa/t, they also have a pyrimidine rich initiator element at the start site (46). The TATA box in conjunction with the cap sequence position the start site of transcription (56).

b) TATA-less promoters

TATA-less promoters exist for housekeeping genes and developmentally regulated genes such as the homeotic genes active during development. The fact that most TATA-less promoters still require all of the general transcription factors, including TFIID, suggests that, with the exception of the TATA

recognition event, the overall initiation mechanism may not be that different from that which occurs at TATA-containing promoters. Many genes that lack TATA and CAAT boxes at the usual positions have multiple transcription initiation sites (57-60). TATA-less promoters can be divided into two classes: those that utilize a single start site (SSS) and those that initiate at multiple start sites (MSS).

The promoters of many genes containing a single start site have an alternative promoter, the initiator (Inr), which encompasses the transcription start site PyPy(C)A_{+1}NT/APyPy and is responsible for localizing the preinitiation complex (61). Drug-sensitive parental Chinese hamster lung cells transcribe pgp1 (P-glycoprotein) from the single +1 Inr site. Multidrug resistant subclones of this cell line utilize additional downstream sites (+32, +42, +52 and +67) (61,62). The regulation of multiple start sites is not clearly understood, but a downstream protein binding sequence, GCTCCC/G, named MED-1 (Multiple start site Element Downstream), was identified in the majority of MSS promoters analysed (63). Ince *et al.* propose that GCTCCC is a conserved element defining a new subclass of RNA polymerase II promoters (63).

1.1.3.2 Assembly of the preinitiation complex

The first step in the assembly of the preinitiation complex is the binding of TFIID to the TATA element (64,65) (Figure 1-2A). TFIID is a multisubunit transcription factor; one of the polypeptide chains, which make this multisubunit factor, is responsible for the recognition of the TATA element, thus aptly termed the TATA-element binding protein (TBP). TFIID contains at least 12 other polypeptides, TBP-associated factors (TAFs) that are also involved in the initial recruitment of TFIID to the DNA but play a more significant role in positive and negative regulatory functions (66).

Once TFIID has identified and bound the core promoter it recruits TFIIA. This stabilises the complex and creates a TFIIA-TFIID-DNA platform (Figure 1-2B) (67). TFIIB and TFIIH join the complex on the TATA box (Figure 1-2C). TFIIF associates with RNA polymerase II. Photocrosslinking studies have localized

TFIIE to the same part of the promoter as TFIIF (68). With the help of TFIIB, TFIIF targets RNA polymerase II to the promoter (Figure 1-2D) (69).

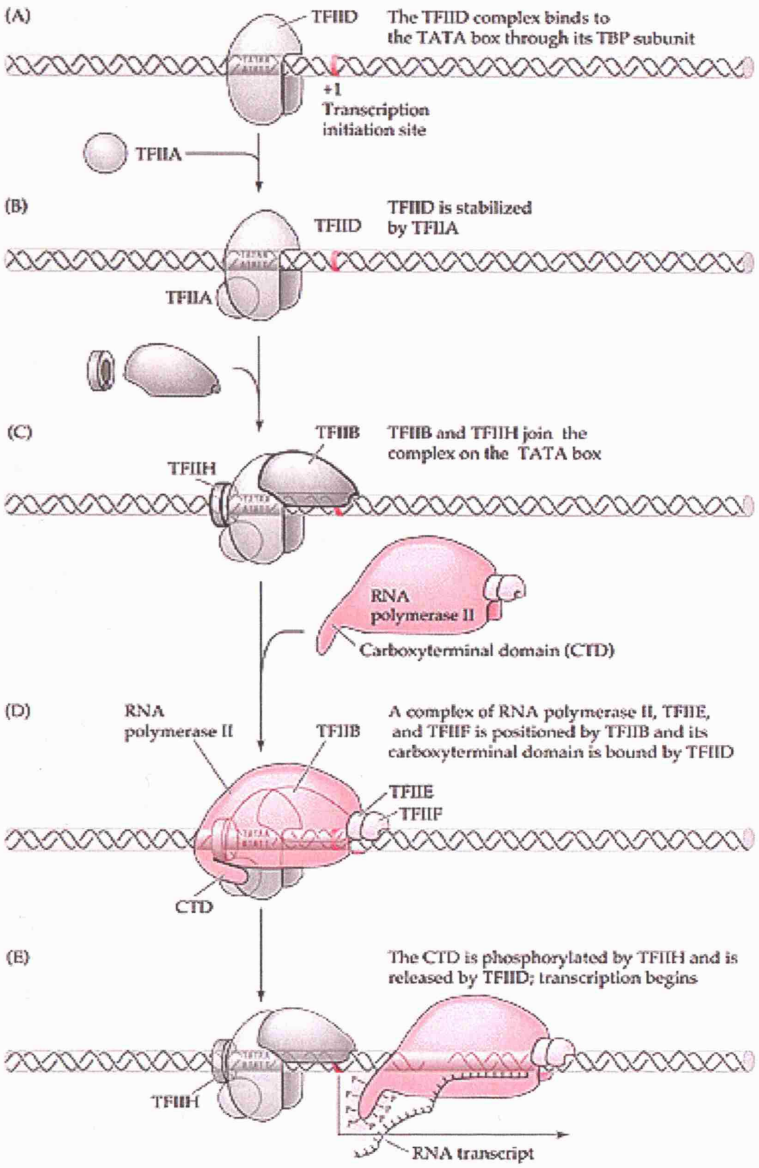

Figure 1-2 The formation of the active eukaryotic initiation complex. (A) The first step in the assembly of the preinitiation complex is the binding of TFIID to the TATA element. **(B)** TFIID then recruits TFIIA. This stabilises the complex and creates a TFIIA-TFIID-DNA platform. **(C)** TFIIB and TFIIH join the complex on the TATA box. **(D)** TFIIF associates with RNA polymerase II and targets it to the promoter. **(E)** Once the preinitiation complex assembly is complete, the helicase activity of TFIIH separates the DNA strands at the transcription start site, which creates an open complex. The kinase activity of the TFIIH subunit then phosphorylates the C-terminal domain of polymerase II allowing it to be released from the promoter and initiate transcription Figure taken from Buratowski *et al.* (67).

TFIIH is unlike any of the other general initiation factors as it possesses catalytic activities, some of which include DNA-dependent ATPase, ATP-dependent DNA helicase and protein kinase capable of phosphorylating the C-terminal domain of the large subunit of RNA polymerase II (70).

More recently scientists have been successful in purifying a number of multiprotein complexes containing RNA polymerase II and a number of the other transcription factors but have not been so successful in separating TFIID and TFIIB which suggests that a "pol II holoenzyme" may form prior to the binding of its recruitment to the TFIID-DNA-TFIIB complex (71).

Once the preinitiation complex assembly is complete, the helicase activity of TFIIH separates the DNA strands at the transcription start site, which creates an open complex. The kinase activity of the TFIIH subunit then phosphorylates the C-terminal domain of polymerase II allowing it to be released from the promoter and initiate transcription (Figure 1-2E) (67,72).

During elongation, *in vitro*, TFIID can remain bound to the core promoter supporting reinitiation of transcription by polymerase II and other general initiation factors (73). Following termination a phosphatase recycles polymerase II to its nonphosphorylated form, allowing the enzyme to reinitiate transcription *in vitro*. This is thought to be a plausible scenario *in vivo*. The binding of the TBP to the core promoter is an essentially slow step, therefore the stable association of TFIID to the core promoter between successive rounds of transcription would make up for the length of time taken for chromatin remodeling and 'pol II holoenzyme' re-formation (74).

Some of the TATA box binding protein-associated factors (TAFs), belonging to the multisubunit complex of TFIID, have been shown to bind activation domains of the activators directly. This interaction has been shown to be essential for activator dependent transcription, but at present very little is known about the specific functions of TAFs in biological processes involving transcriptional regulation (75). The three-dimensional structure of TFIID is said to be saddle shaped. This is very appropriate for its role in both DNA and transcription factor

binding. The concave underside of the saddle supports the DNA binding while the seat of the saddle binds various components of transcriptional machinery (76).

1.1.3.3 Cis/Trans acting regulatory elements

Promoters do not work on their own, *cis*-acting elements, including enhancers and locus control regions (LCR), and *trans*-acting proteins, such as transcription factors, and activators, enhance transcription in specific cell types. *Cis*-acting factors are defined by DNA elements that act only on the sequence of DNA that they are physically attached. Trans-acting factors are diffusible products, meaning their ultimate function is not dependent on their location in the genome once they are expressed. These different factors all work together to regulate the transcription of genes (Figure 1-3).

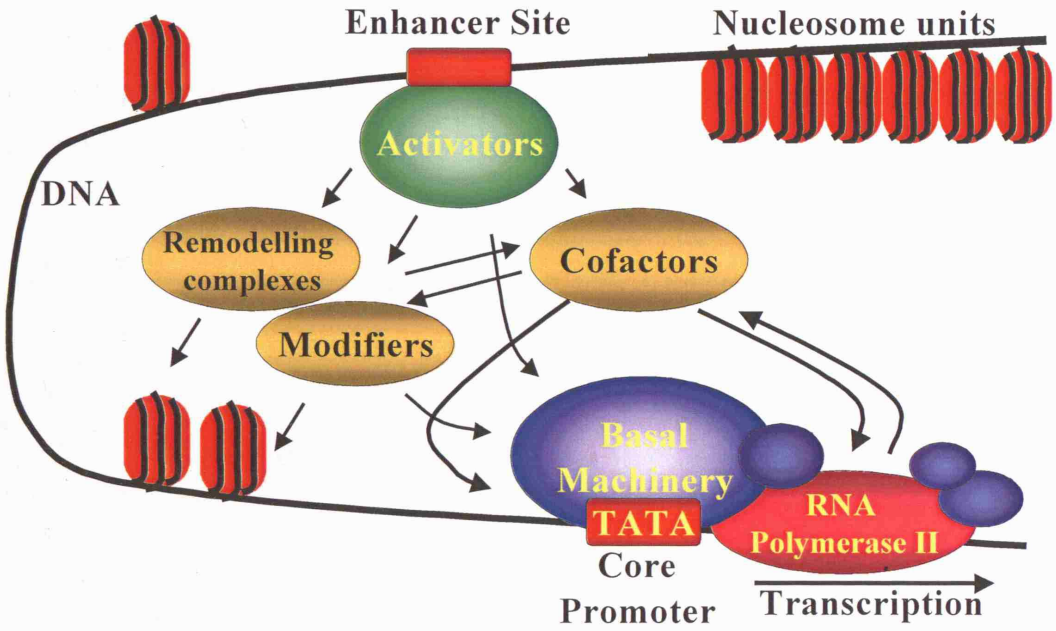

Figure 1-3 Schematic overview of the interaction between several classes of proteins that cooperate to initiate transcription of DNA. The chromatin remodeling complexes and histone tail modifiers destabilize the nucleosome in order to make the DNA accessible. Activators recruit the chromatin remodeling complexes, modifiers, cofactors and basal transcription factors in order to access the DNA and assemble the pre-initiation complex on the core promoter. Many cofactors are required to mediate signals between transcriptional activators and the basal transcription machinery (Figure adapted from Lodish *et al.*(10))

a) Enhancers

Enhancers activate transcription from a much further distance than promoters and are usually 100-200 base pairs long. Although enhancers often lie within a few kilo-bases of the cap site, in some cases they lie much further upstream or downstream from the cap site or within an intron. In addition, some genes are controlled by more than one enhancer region.

Many enhancers, in the absence of a promoter, can stimulate transcription of a reporter gene when transfected into cell lines. In an *in vivo* context, circumstances are different, and expression of genes requires additional regulation.

Some transcription factors ("Enhancer-binding protein") bind to regions of DNA, the enhancer, thousands of base pairs away from the gene they control. Binding increases the rate of transcription of the gene. The detailed mechanism of enhancer function is not understood. One possibility, shown by Botchan *et al.* (77), is that enhancer-binding proteins, in addition to their DNA-binding site, have sites that bind to transcription factors assembled at the promoter of the gene that enables them to draw the DNA into a loop (77).

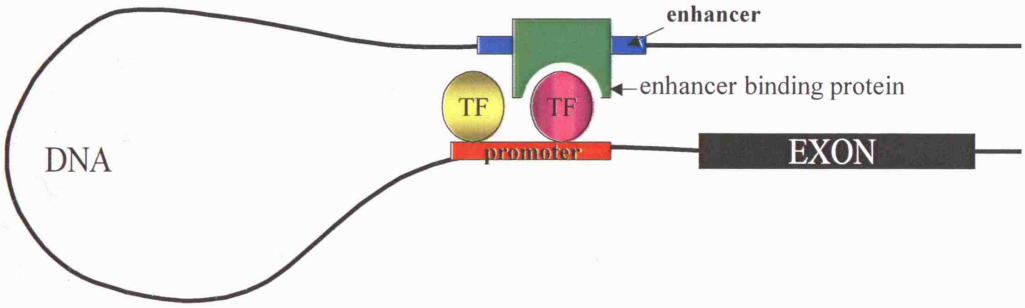

Figure 1-4 A schematic representation of an enhancer binding protein recruiting the enhancer, thousands of base pairs away, to the promoter via its interactions with transcription factors drawing the DNA into a loop structure. Botchan *et al.* created an artificial DNA molecule containing an enhancer and a transcription factor binding site. Under an electron microscope the DNA was seen in loops. This showed that enhancer-binding proteins, in addition to their DNA-binding site, have sites that bind to transcription factors assembled at the promoter of the gene that enables them to draw the DNA into a loop (Figure adapted from Botchan *et al.* (77)).

Botchan *et al.* created an artificial DNA molecule with several promoter sites for Sp1 about 300 bases from the 3' end of the molecule (77). Sp1 is a zinc-finger transcription factor that binds to the sequence 5' GGGCGG 3' found in the

promoters of many genes, especially "housekeeping" genes. This artificial DNA had several enhancer sites, which bind the enhancer-binding protein E2, about 800 bp from the 5' end of the molecule and 1800 bases from the SP1 binding site. When these DNA molecules were added to a mixture of Sp1 and E2, the electron microscope showed that the DNA was drawn into loops with "tails" of approximately 300 and 800 base pairs (77).

b) Locus control regions (LCR)

LCRs are areas of DNA containing a cluster of DNase I hypersensitive sites. They are thought to be able to confer transcriptional competence onto a gene regardless of their position on the genome, (78). Recent evidence suggests enhancers and LCRs increase the probability at which a pre-initiation complex (PIC) is formed in any given cell rather than increase the frequency at which transcription is initiated (79,80).

Proximal elements, such as those controlling the HSV tk gene, are usually found between 50 and 200 bp upstream of the transcriptional start site (81). These elements lose their influence when moved an additional 15-20 bp further from the promoter. Whereas distal enhancer elements can be found a long way from here, for example the regulatory element of IL4 cytokine that is located ~10kb upstream of the transcription start site of IL4. This distant regulatory element also drives expression of a second cytokine gene, called IL5, that is located ~120kb away (82). Active regulatory elements can be located hundreds of kilo-bases away from the genes that they control, especially in regions of low gene density.

c) CCAAT box

The CAAT box - consensus sequence CCAAT - is also often found between 40 and 100 bp upstream of the start point of transcription. The transcription factor CTF or NF1 binds to the CAAT box.

d) GC box

The GC box is a common element in eukaryotic class II promoters. Its consensus sequence is GGGCGG. It may be present in one or more copies, which can be

located between 40 and 100 bp upstream of the start point of transcription. The transcription factor Sp1 binds to the GC box.

e) Transcription Factors

Transcription factors vary in their structures but they all have two functional domains.

a) The DNA-binding domain, which recognizes the specific DNA sequence of the enhancer, and binds to it.
b) The transactivation domain, which interacts with other proteins and thereby, increases the rate of transcription.

Transcription factors can be classified according to the structure of their DNA binding domain. There are four basic types: Zinc finger proteins, helix-turn-helix proteins, leucine zipper proteins, helix-loop-helix proteins; all named based on distinct properties present in these proteins (12).

1.1.4 Transcript Processing and Modification

Before a gene transcript can be transported out of the nucleus it has to undergo three processing events. RNA processing involves addition of a 5' cap, addition of a 3' poly (A) tail, and removal of introns. Although these reactions are biochemically different processes, they are linked by their requirement for the C terminal domain (CTD) of RNA polymerase II, thereby influencing the specificity and efficiency of each other. The CTD of vertebrate RNA polymerase II is composed of 52 heptad repeats. It is required for transcriptional activation and repression, efficient capping, splicing and polyadenylation of mRNA transcripts (83).

In recent years evidence of the tight linkage of splicing and poly A addition to transcription, and the influence that they have on each other have made scientists conclude that most mRNA processing reactions occur co-transcriptionally (84). These processing events represent another level of regulation of gene expression, particularly with regard to splicing out of introns. Regulation at this level determines whether RNA gets processed and/or which exons are retained in the

mRNA. If RNA is not processed, it will not be transported out of the nucleus, and will not be translated (46).

1.1.4.1 End modification

a) 5' capping

Transcripts made by RNA polymerase II are distinguished from those made by polymerase I and III by a $m^7G(5')ppp(5')N_1$ cap. The 5'-ends of all eukaryotic pre-mRNAs studied thus far are converted to cap structures. The cap is thought to influence splicing of the first intron, and is bound by 'cap-binding' proteins, CBP80 and CBP20, in the nucleus (84). It is important for translation initiation, and it also interacts with the poly(A)terminus, via proteins, resulting in circularisation of the mRNA to facilitate multiple rounds of translation. The cap is also important for mRNA stability, protecting it from 5' to 3' nucleases, and is required for mRNA export to the cytoplasm (85). The capping reaction usually occurs very rapidly on new transcripts, after the synthesis of only a few nucleotides by RNA polymerase II. The reaction involves the conversion of the 5'-end of the growing transcript from a triphosphate to a diphosphate by a RNA 5'-triphosphatase, followed by the addition of a guanosine monophosphate by the mRNA guanylyltransferase to form a 5'-5'-triphosphate linkage. This cap is then methylated by 2'-O-methyltransferases.

The capping enzymes are always in the right place at the right time because they are recruited by the C-terminal domain of RNA polymerase II when it is phosphorylated by TFIIH, close to the promoter (86). Within the production of the first 500 bases of the pre-mRNA, the guanylyltransferase is released, whereas the methyltransferase release more slowly and a significant amount remains associated with the RNA polymerase right to the 3'end. Between the two enzymes they manipulate polymerase II function at the 5'end of a gene to achieve co-transcriptional capping. It has been suggested that they also act as checkpoints to hold up elongation until the addition of the cap (84).

This initial cap structure is then recognised by the cap binding complex (CBC), containing CBP20 and CBP80 proteins. Once the RNA has been transported through the nuclear pore complex, the nuclear cap binding proteins are replaced

by the 4E subunit of the cytoplasmic translation initiation factor, eIF4F. The 4E protein binds directly to the mRNA cap and promotes ribosome attachment. Pre transport into the cytoplasm the CBC promotes efficient interaction between U1 snRNP and the 5' splice site (83). The interaction between U1 snRNP and the 5'splice site is one of the earliest steps in spliceosome assembly (described in more detail in 1.1.4.2.1).

Turnover of mRNA is an important factor in the regulation of gene expression and the rate of de-capping is a primary determinant of the half-life of mRNA. Any mRNA lacking a cap structure is rapidly degraded by 5'→ 3' exonucleases (84). Furthermore, in the cytoplasm, the cap binds translation initiation factors which promote the interaction of the ribosomal subunits with the mRNA thus enhancing translation. The efficiency of translation is further augmented by the interaction of eIF4G with poly A binding protein I, PABPI, resulting in a translationally competent circular mRNA protein complex. The interaction between PABP and the initiation factors associated with the 5' terminus of an mRNA may also explain the role of PABP during mRNA turnover, as it protects the 5' cap from attack by the decapping enzymes (87).

b) 3' poly A addition

The 5' cap and 3' poly A tail are both formed in the nucleus and serve a similar purpose in the cytoplasm but there are big differences. The main difference is the complexity in the protein machinery required for 3' end formation, specifically for polyadenylation. More than a dozen polypeptides are required, most of whose function remains unknown (83).

The upstream fragment generated by 3' cleavage of the pre-mRNA receives a poly(A) tail of approximately 250 adenine residues in a reaction depending on the AAUAAA sequence 10 to 30 nucleotides upstream of the 3' end. Polyadenylation in vertebrates requires two main cis-acting elements that straddle the cleavage site, identified by the AAUAA hexamer 10-30 nucleotides upstream and a GU-rich element downstream. Cleavage and polyadenylation specificity factor, CPSF, recognizes and binds to the pre-mRNA at the AAUAAA hexamer, while cleavage stimulatory factor, CstF, interacts with the

U-rich or GU-rich region enhancing the affinity of CPSF for the hexamer. Additional cleavage factors, CFI and CFII are also involved in the cleavage step (88). Assembly and cleavage are followed by polyadenylation. Stimulation of poly-(A) polymerase, PAP, by both proteins, CPSF and CstF, results in the elongation of the RNA. CPSF recruits PAP to the AAUAAA sequence and forms a complex adequate for slow polymerization of the poly-A-tail. Poly-(A)-binding protein (PABP), which binds the growing poly-(A)-tail once it has reached a length of about ten nucleotides, catalyses the addition of the adenosine residues until the length of the tail reaches about 200-250 nucleotides. PABP is essential to terminate elongation but the mechanism is not understood (84).

1.1.4.2 Splicing and occasional RNA editing

Pre-mRNA splicing, the removal of non-coding introns, is a necessity for the expression of most eukaryotic genes. This essential reaction proceeds in two steps. Step 1 is the 5' splice site cleavage and ligation of the introns 5' end to the branch site. Step 2 is the 3' splice site cleavage and ligation of the 5' and 3' exon.

A number of studies have shown that splicing occurs as mRNA transcription occurs. The splicing of introns is a complex reaction mediated by a complex composed of RNA and protein molecules, the spliceosome. The spliceosome catalyses the reaction between the 2' hydroxyl group of an adenine, and the 5' phosphate end of the intron, creating a lariat loop (Figure 1-5F). Only RNA has 2' OH groups, thus preventing splicing occurring in DNA. The lariat reaction produces a 3' OH group on one exon, enabling it to join the 5' phosphate of the joining exon (89).

Sequences that are essential for the removal of introns are based at the boarders of the introns and exons. In higher eukaryotes the 5' splice site sequence is AG|**GU**RAGU (the splice site is represented by a vertical bold line). The constant nucleotides are underlined and in bold (R = purine and Y = pyrimidine). The sequence of the 3' splice site is Y**AG**| and is preceded by a stretch of pyrimidine residues. There is one other important site involved in the splicing mechanism,

the branch site. This site is usually located between 18 and 40 nucleotides upstream of the 3' slice site and is recognised by the sequence YNCUR**A**Y (where N is any nucleotide, the site of the branch formation is shown in bold and underlined) (89,90).

1.1.4.2.1 Pre mRNA splicing

Splicing of precursor mRNA takes place in the spliceosome. At each round of processing a new large RNA/protein complex assembles on the RNA. The spliceosomal machinery is composed of five uridine rich small nuclear ribonucleoprotein (snRNPs) transiently associated to more than 760 non-snRNP splicing factors (RNA helicases, SR splicing factors etc.) (91).

Pre-mRNA transcripts become rapidly associated with many RNA-binding proteins, including hnRNP proteins, cap-binding proteins and SR proteins. The E complex, or early complex, is the first intermediate to be detected in spliceosome assembly in vitro. When a functional 5' splice site is located, it is bound by the U1 snRNP. The splicing factor U2AF (65 and 35 kDa subunits) binds to the polypyrimidine tract, (Y)n, and the AG nucleotides at the 3' splice site, respectively (Figure 1-5A). SF1/mBBP binds to the branch site. The SR proteins and U2AF interact with each other, facilitating their binding to the pre-mRNA. Next, the E complex is converted to the A complex, in the presence of ATP (89,90,92).

The A complex is the first ATP-dependent complex in spliceosome assembly. U2AF recruits the U2 snRNP to bind to the branch site in the E complex in an ATP-dependent fashion, to form the A complex (Figure 1-5B). The U2 snRNA binds to the branch site, causing the branch-site adenosine to bulge out, which puts it in the correct position for nucleophilic attack at the 5' splice site. The A complex acts as a substrate for formation of the B complex (89,93).

The formation of the B complex is ATP-dependent, and both the 5' and 3' splice sites are essential for B complex assembly. The U4 and U6 snRNPS are joined, and this U4:U6 complex associates with the U5 snRNP to form a tri-snRNP

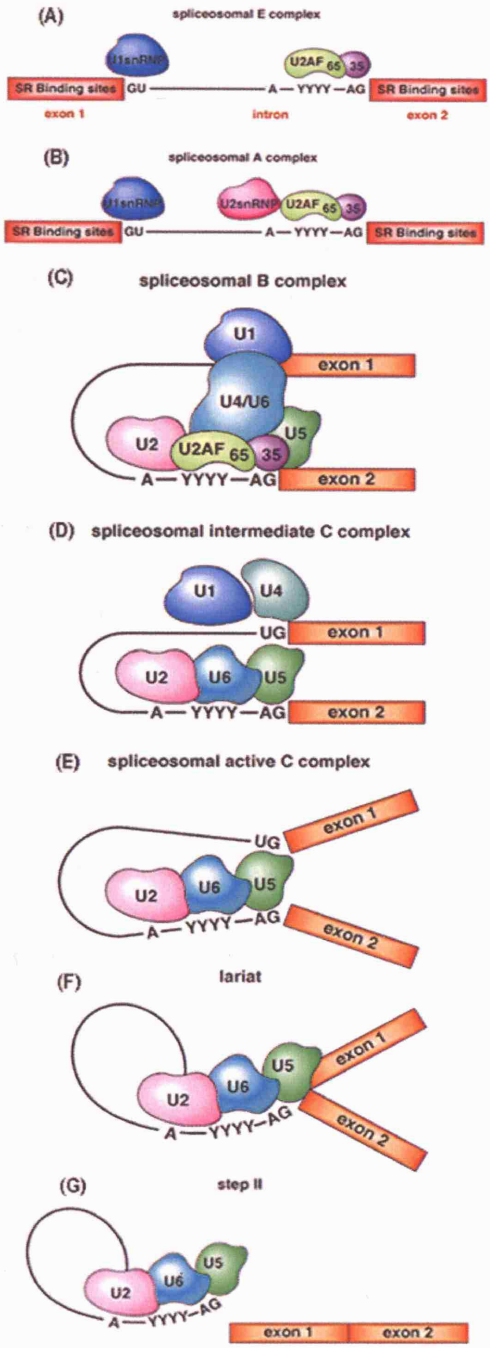

Figure 1-5 Complex formation along the splicing pathway. (A) The splicing factor U2AF (65 and 35 kDa subunits) binds to the polypyrimidine tract, (Y)n, and the AG nucleotides at the 3' splice site, respectively. **(B)** The A complex is the first ATP-dependent complex in spliceosome assembly. U2AF recruits the U2 snRNP to bind to the branch site in the E complex in an ATP-dependent fashion, to form the A complex. **(C)** U4:U6 complex associates with the U5 snRNP to form a tri-snRNP particle. This tri-snRNP particle then binds to the spliceosomal A complex, to form the spliceosomal B complex **(D)** The splicing intermediates are rapidly converted to splicing products. The spliced products are released very rapidly. **(E)** The active C complex is formed due to a conformational change in the intermediate C complex **(F)** The bond between the 5' exon and the 5' end of the intron is cleaved and a new bond between the 5' end of the intron and the adenine nucleotide at the branch site is formed. This results is a lariat-shaped intermediate, with the intron still attached to the 3' exon. **(G)** Cleavage of the transcript at the 3'splice site, and results in ligation of the two exons and excision of the intron (Figure adapted from Rappsilber *et al*. (90).

particle. This tri-snRNP particle then binds to the spliceosomal A complex, to form the spliceosomal B complex (Figure 1-5C) (89,90,92).

The spliceosomal C complex is a very transient intermediate; the splicing intermediates are rapidly converted to splicing products. Also, the spliced products are released very rapidly. No complex containing both the splicing products has been isolated. Conversion of the spliceosomal B complex to the spliceosomal C complex requires ATP (Figure 1-5D) (92,94).

The active C complex is formed due to a conformational change in the intermediate C complex (Figure 1-5E). After formation of the active C complex, the splicing reactions occur very rapidly (89,90).

In the first catalytic step of mRNA splicing, the bond between the 5' exon and the 5' end of the intron is cleaved and a new bond between the 5' end of the intron and the adenine nucleotide at the branch site is formed. This results is a lariat-shaped intermediate, with the intron still attached to the 3' exon (Figure 1-5F) (90).

The second step of the splicing reaction results in cleavage of the transcript at the 3'splice site, and results in ligation of the two exons and excision of the intron. The liberated intron remains associated with U2, U5 and U6 snRNPs and the spliced RNA is released from the spliceosome (Figure 1-5G) (89,90).

1.1.4.2.2 Alternative splicing

Alternative splicing of pre-mRNA is a common mechanism of regulating gene expression. It is thought that at least half of all genes in the human genome undergo alternative splicing. Alternative splicing is controlled by the binding of trans-acting factors to cis-acting sequences in the pre-mRNA resulting in the spliceosome being directed toward or away from certain splice sites depending on the sequences being enhancer or suppressor elements (95).

Alternative splice site recognition is mediated by splicing factors belonging to the SR (serine rich) protein family. This family of proteins has a highly conserved amino terminal RNA recognition motif (RRM) and a carboxyl terminal domain that consists, almost exclusively, of serine and arginine residues (RS domain) (96). There are two alternative methods of splice site selection, arginine-serine (RS) domain-dependent (Figure 1-6A) and RS domain-independent (Figure 1-6B) (97).

Exonic splicing enhancer (ESE) and exonic splicing silencer (ESS) elements are present in most constitutive and alternative exons. The RNA binding domain of the SR proteins recognizes and binds to the ESE and promotes splicing by recruiting members of the spliceosome complex. The interaction of U2 and U1 at the 3' and 5' splice sites is then strengthened via the arginine-serine rich domain (RS) of the SR protein (Figure 1-6A) (97).

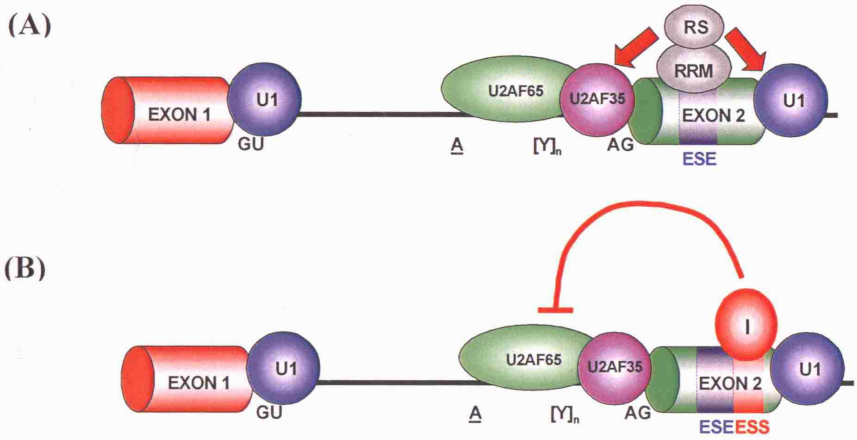

Figure 1-6 RS-domain dependent and independent action of SR proteins. (A) The SR protein binds to the Exonic Splicing Enhancer (ESE) and through its interactions, via its RS domain, strengthens the interaction of U2AF35 and U1snRNP at the 3' and 5' splice sites of the pre-mRNA respectively. **(B)** The splicing inhibitor protein binds to the Exonic Splicing Silencer (ESS) and inhibits the binding of U2AF65 and U2AF35 to the 3' splice site (Figure adapted from Cartegni *et al.*(97)).

RS domain-independent action of SR proteins is seen when the protein overcomes the effect of splicing inhibitor proteins. The splicing inhibitor protein binds to the ESS and inhibits the binding of the U2 members of the spliceosome at the 3' splice site. However, successful binding of the SR protein to the ESE overcomes the effect of the inhibitor and allows proper binding of the spliceosome components.

There are seven basic types of alternative splicing (Figure 1-7) which result in, for example, the addition or removal of protein domains, an altered protein sequence due to the shifting of the reading frame, a truncated protein via the introduction of a stop codon, or the insertion or deletion of cis regulatory elements via alternative splicing of 5' and 3' untranslated regions having an effect on the localization, stability or translation efficiency of the mRNA (90).

Examples of all of these types of splicing have been identified in the human genetic repertoire. The alternative inclusion of a cassette exon is very common (Figure 1-7a). An example of an insertion of an exon cassette occurs in the c-src protooncogene generating an extra SH3 protein–protein interaction domain that alters its binding to other proteins (95).

The mutually exclusive use of alternate exons (Figure 1-7b) is evident in the fibroblast growth factor receptor 2. The use of one exon produces a receptor with a high affinity to one particular growth factor, whereas the use of the other exon produces a high affinity receptor to a different ligand (95).

Wt1, the Wilms' tumour suppressor gene is regulated via alternative 5' splice sites (Figure 1-7c). The different forms play distinct role in gonad and kidney formation (95,98).

Female somatic sexual differentiation in *Drosophila* is regulated via the *transformer (tra)* gene. The single long open reading frame of *tra* produced in the female gives rise to a functional protein where as the shorter version made in the male does not make any protein (Figure 1-7d) (95,98).

Calcium homeostasis in the thyroid gland and vasodilatation in the nervous system are both controlled by a hormone/neuropeptide transcribed from the same gene. Alternative terminal exons decide the fate of the transcript (Figure 1-7e) (95).

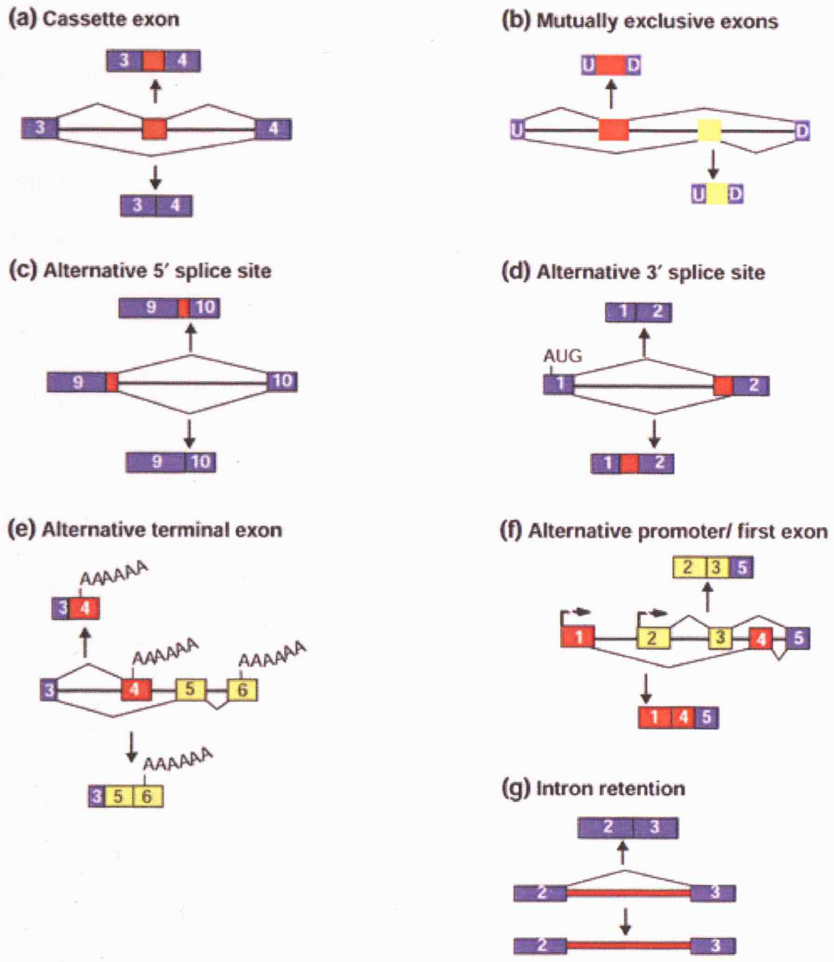

Figure 1-7 The seven basic types of alternative splicing. **(a)** An example of an insertion of an exon cassette occurs in the c-src protooncogene generating an extra SH3 protein–protein interaction domain that alters its binding to other proteins. **(b)** The mutually exclusive use of alternate exons is evident in the fibroblast growth factor receptor 2. The use of one exon produces a receptor with a high affinity to one particular growth factor, whereas the use of the other exon produces a high affinity receptor to a different ligand. **(c)** Wt1, the Wilms' tumour suppressor gene is regulated via alternative 5' splice sites. The different forms play distinct role in gonad and kidney formation. **(d)** Female somatic sexual differentiation in *Drosophila* is regulated via the *transformer (tra*) gene. The single long open reading frame of *tra* produced in the female gives rise to a functional protein where as the shorter version made in the male does not make any protein. **(e)** Calcium homeostasis in the thyroid gland and vasodilatation in the nervous system are both controlled by a hormone/neuropeptide transcribed from the same gene. Alternative terminal exons decide the fate of the transcript. **(f)** Differential promoter usage of the myosin light chain gene leads to different first exons which splice on to unique downstream exons resulting in two distinct protein isoforms, MLC1 and MLC3. **(g)** The rarest form of alternative splicing is that of intron retention. Under these circumstances an intron is left in the mature RNA exported to the cytoplasm. This occurs in patients suffering from muscular dystrophy. Figure from Ladd *et al*. (95).

Alternative promoter usage is a very common mode of gene regulation. The most extensively studied example of this type of alternative splicing is that of the myosin light chain gene (MLC). Differential promoter usage leads to different

first exons which splice on to unique downstream exons resulting in two distinct protein isoforms, MLC1 and MLC3 (Figure 1-7f) (95,98).

The rarest form of alternative splicing is that of intron retention. Under these circumstances an intron is left in the mature RNA exported to the cytoplasm. This occurs in patients suffering from muscular dystrophy. The retention of the second intron introduces a premature stop codon leading to the down regulation of the protein resulting in problems in muscle relaxation (Figure 1-7g) (95).

1.1.4.2.3 mRNA proofreading

Proofreading, post DNA replication and RNA transcription, have been well studied, but these types of mechanisms also exist following RNA splicing to assess the position of the stop codon. If the stop codon is detected anywhere other than in the last exon then processes are initiated to either degrade the transcript or remove the premature stop codon. If the stop codon is missing the mRNA is also targeted for destruction (99).

Some inherited diseases are caused by the insertion of premature stop codons. For example genomic instability in tumours often results from the insertion of a premature stop codon. The presence of these premature stop codons can induce alternative splicing which could provide disease and tumour-specific diagnostic markers (100).

1.1.4.3 Alternative splicing, multiple promoters and chromatin structure direct temporal and tissue-specific gene expression

Development of organisms, as complex as eukaryotes, requires the transcription of over 20-25,000 genes in a temporal and tissue-specific pattern. One of the key problems in the field is in understanding how an organism can achieve such diversity. Several genes have been identified that undergo tissue-specific patterns of alternative splicing. This generates biologically different proteins from the same gene.

1.1.4.3.1 N-terminal variant isoforms

One method employed is the use of multiple promoters. Heterogeneity in the 5' region of the gene is a common feature among steroid/retinoic acid receptors. Some members have functionally distinct isoforms arising from the use of different promoters and/or alternative splicing, two such examples of proteins with N-terminal variant isoforms are the:

a) Progesterone receptor

There are two forms of the human progesterone receptor (hPR), A and B, which originate from alternative ATG translation start sites, ATG1 +744 and ATG2 +1236 (Figure 1-8A) (101). Work done by Kastner *et al.* demonstrated that the A and B forms are encoded by separate RNAs which are transcribed from two distinct promoters, both of which are under estrogen control. Progesterone responsive genes are activated differentially by the different forms of hPR (101).

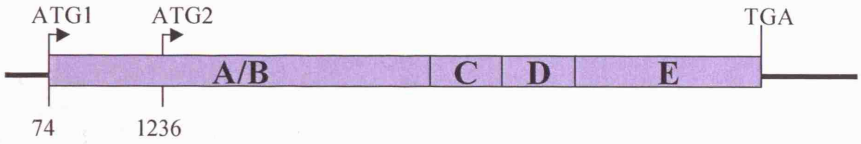

Figure 1-8 The human progesterone receptor gene. There are two forms of the human progesterone receptor (hPR), A and B, which originate from alternative ATG translation start sites, ATG1 +744 and ATG2 +1236. The N-terminal region A/B contains the two ATG translation start sites of hPR forms A and B (adapted from Kastner *et al.* (101)).

b) Thyroid receptor (TR)

Regulation of TR gene expression is more complex that that of the PR. There are five different TR mRNA transcripts made from three alternate promoters and alternative splicing. The mRNA encoding each isoform of the thyroid receptor displays characteristic patterns of developmental, tissue-specific and hormonal regulation (50).

1.1.4.3.2 Variable 5'UTR but identical protein

Tissue-specific alternative promoter usage generates multiple transcripts of the human (102) and mouse oestrogen receptor α (ERα) (48), the human (103) and rat (104) mineralocorticoid receptors and the mouse glucocorticoid receptor (47),

which all differ in their 5' untranslated regions (UTR) but encode identical proteins.

a) Mouse estrogen receptor α (ERα)

Studies have shown that the mouse ERα gene is transcribed from at least five different promoters, all of which have a high degree of homology between human, rat and mouse. Work done by Kos *et al.* revealed not only a tissue-specific but also a sex-specific expression pattern of all of the variants. Three out of the five mRNAs observed were in all tissues tested but are expressed at different levels, one was expressed everywhere except the aorta and lung. The most diverse mRNA transcript was expressed only in the liver and at a three-fold higher level in the female liver than in the male liver (48). Each cell type possesses a different composition of transcription factors and regulatory proteins that selectively direct the utilization of particular promoters.

1.1.4.3.3 Multiple promoters that direct both tissue-specific expression and the production of N-terminal variants of the protein

a) Human vitamin D receptor

One specific hormonal form of vitamin D plays a vital role in calcium and phosphate homeostasis and is involved in the regulation of cell growth and differentiation in specific target tissues (105). The human vitamin D receptor, hVDR, has multiple promoters that direct both the tissue-specific expression of the gene and the production of N-terminal variants of the protein (Figure 1-.9).

Transcription initiation from two of the untranslated exons, 1a and 1f, 9 kilo-bases (kb) apart, generates tissue-specific transcripts that differ in their 5'UTR but encode the same 427 amino-acid protein. Transcripts made from the distal exon, 1f, were found only in kidney tissue and one particular intestinal cell line. These transcripts were undetectable in other intestinal cell lines suggesting that expression of exon-1f transcripts is cell type-specific.

Transcription from a third untranslated exon and subsequent alternative splicing generates transcripts able to encode proteins with additional amino-acids at the N terminus (Figure 1-9) (106).

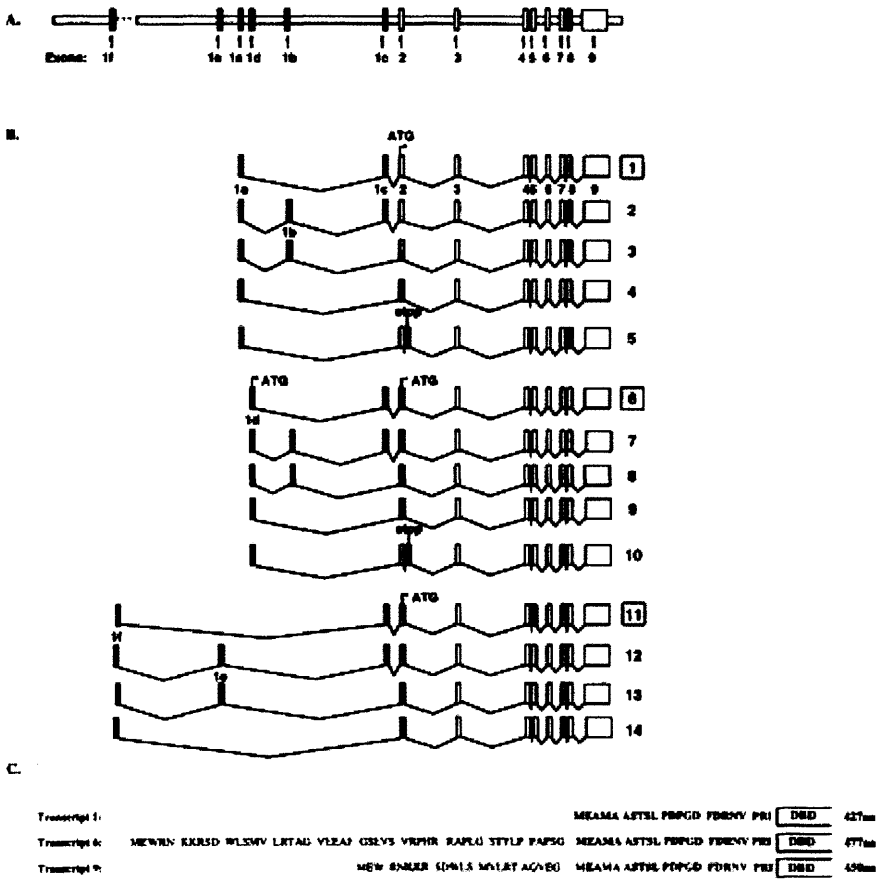

Figure 1-9(*A*) Human VDR gene locus (*B*) Structure of hVDR transcripts. Transcripts 1-5 originate from exon 1a. Transcripts 6-10 originate from exon 1d and transcripts 11-14 originate from exon 1f. Boxed numbers indicate the major transcript. While all transcripts have a translation initiation codon in exon 2, exon 1d transcripts have the potential to initiate translation upstream in exon 1d, with transcripts 6 and 9 encoding VDR proteins with extended N termini. **(*C*) N-terminal variant proteins encoded by novel hVDR transcripts.** Transcript 1 encodes the 427-aa hVDR protein. Transcripts 6 and 9 code for a protein with an extra 50 aa or 23 aa, respectively, at the N-terminal. Figure from Crofts *et al.*(106).

1.1.4.3.4 Other well-characterized examples of tissue-selective expression determined by multiple promoters

a) Lck

lck is a well characterized gene that encodes a protein tyrosine kinase that plays a key role in signaling mediated through the T cell receptor. The expression of the *lck* gene is restricted to lymphoid cells, and occurs predominantly, but not exclusively in T cells. Studies on the regulation of *lck* gene expression have

identified two independent promoter elements, the proximal and distal promoters (107,108). The proximal promoter is used in thymocytes, but not in peripheral T lymphocytes. The distal promoter is active in all stages of T cell development but has a higher activity in more mature T cells (109,110).

b) Protein Kinase C- η (PKCη)

PKC-η is a member of the PKC family, a large group of phospholipid dependent serine/threonine protein kinases involved in a number of biological processes in the cell, including cell proliferation and differentiation. Different isoforms of PKC show distinct patterns of tissue distribution suggesting that the different isoforms are regulated independently and may have independent functions (111). PKC-η is predominantly expressed in epithelial tissue, such as skin, lung and intestine, and in the skin is limited to keratinocytes in the upper layers of the epidermis (112). Two major transcription start sites were identified which are only 12 nucleotides apart, this prompted further studies which demonstrated that the upstream region of the PKC-η gene contains enhancer and silencer regulatory elements sufficient to drive cell type-specific transcription (113).

1.1.4.3.5 Extraordinary examples of gene regulation via the 5'UTR

a) Dystrophin gene

The dystrophin (*DMD)* gene has more than 79 exons spanning about 2.4 mega-bases (Mb) of DNA. At least eight different alternative promoters are used. Four of these are located near the conventional start site and include a brain cortex-specific promoter, a muscle-specific promoter located 100 kb downstream, a promoter which is used in Purkinje cells of the cerebellum and located a further 100 kb downstream, and a lymphocyte-specific promoter (Figure 1-10). The proteins made from these promoters are the largest isoforms with a molecular weight of 427 kDa (referred to as Dp427). The four Dp427 isoforms differ in their extreme N-terminal amino acid sequence as a result of using four different alternatives for exon 1 (Figure 1-10) (114,115).

In addition to the four alternative promoters encoding the conventional large isoforms, at least four other alternative internal promoters can be used.

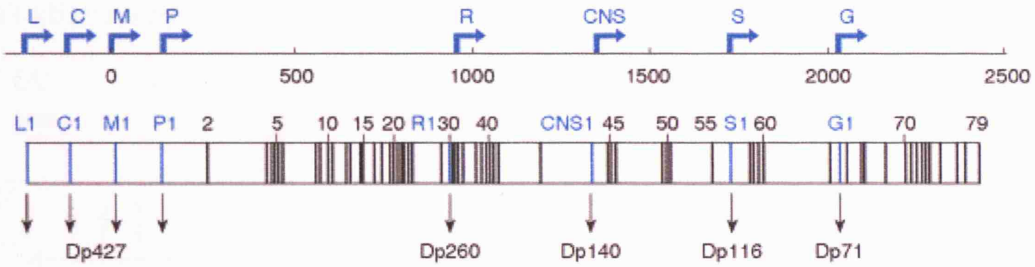

Figure 1-10 **At least eight distinct promoters can be used to generate cell type-specific expression of the dystrophin gene**. The positions of the eight alternative promoters are shown in blue: L, lymphocyte; C, cortical; M, muscle; P, Purkinje; R, retinal; CNS, central nervous system; S, Schwann cell; G, general. Each promoter uses its own first exon (blue vertical line: L1, C1, M1, P1, R1, CNS1, S1 and G1) together with downstream exons (black vertical line). All 78 downstream exons are used in the case of the full-length C-, M- and P-dystrophins which generates a product of about 427 kDa (Dp427). The other promoters are located immediately upstream of indicated exons as follows: R, exon 30; CNS, exon 45; S, exon 56; G, exon 63. The initiation of translation for the Dp140 isoform does not occur until exon 51, although the promoter is thought to be in intron 44 (adapted from Nudel *et al.* (114)).

Transcription from these promoters uses only a downstream subset of the exons, resulting in significantly smaller isoforms: a Dp260 isoform produced in retinal cells; a Dp140 isoform produced by many cells in the brain and kidney; a Dp116 isoform produced in Schwann cells and a small Dp71 isoform produced in many cell types (Figure 1-10). As well as multiple promoter usage generating a number of proteins and tissue specificity, alternative splicing occurs at the 3'end further increasing diversity (115).

b) α-dystrobrevin gene

The 270 kb 5'UTR belonging to the α-dystrobrevin gene is not a common phenomenon. α-dystrobrevin is a phosphoprotein proposed to play a role in synaptic structure and function (116,117). The most distal promoter, active in the brain, is 120 kb from the muscle promoter. 70 kb downstream from the muscle promoter is a third promoter, which is another 80 kb from exon 1. The α-dystrobrevin gene has 25 coding exons spanning 170 kb of genomic DNA yet its seven small-untranslated exons span 270 kb, resulting in the total size of the gene exceeding 400 kb (Figure 1-11) (118). Knock out of the gene in mice results in muscular dystrophy. The gene expression pattern of α-dystrobrevin resembles that of dystrophin where different isoforms are generated via alternative splicing

in both brain and muscle. Promoters at the 5' end of the gene control the different 14 kb transcripts (119,120).

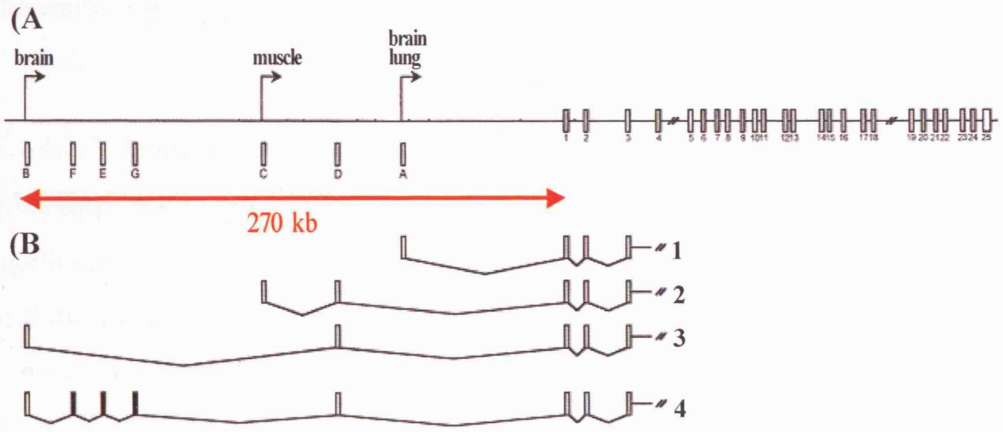

Figure 1-11 Genomic organization and alternative splicing of α-dystrobrevin 5'-UTR exons. **(A)** A schematic representation of the α-dystrobrevin gene. The three promoters (*arrows*) and the approximate position of exons A-G in the first 270 kb of the α-dystrobrevin gene are shown relative to the 25 coding exons 1-25. **(B)** A schematic illustration of the predominant splicing patterns in brain and muscle. The four transcripts are generated by alternative usage of exons A, B, C, and D (*open boxes*) with lines indicating regions that are spliced out of the primary transcripts. In a minority of the third transcripts one or the other of three internal exons, exons F, E, and G (*black boxes*), were found to be spliced between exon B and D and therefore generate further complexity in the 5'-untranslated region. Figure adapted from Holzfeind *et al.* (118).

c) hepatocyte nuclear factor 4α (HNF-4α)

HNF-4α does not have quite such an enormous 5'UTR as the α-dystrobrevin gene but with 45 kb of genomic DNA between its two promoters it unusually large (121). The HNF-4α gene encodes a cell-specific transcription factor that is expressed in liver, kidney, intestine, stomach and pancreas and it has been seen to be involved in early onset diabetes. Transcription from its distal promoter has been identified as the major transcription site used in pancreatic β-cells. A mutation in one particular binding site in this promoter co-segregates with a locus linked to diabetes (121,122).

1.1.4.3.6 Tissue specific expression determined by chromatin structure

Vav, a GTP/GDP exchange factor involved in the regulation of a G protein of the Rho/Rac/CDC42 family (123), is an example of a gene that is regulated in a tissue-specific manner via the organisation of the chromatin. Vav is expressed in all haemopoietic cell types but very few others. Five DNase I hypersensitive sites (HS) were found in hematopoietic cells of three distinct lineages (myeloid,

erythroid and lymphoid), but none were found in fibroblasts. These five sites, which are spread over 14 kb are likely to represent the major regulatory sites for the control of vav gene expression (124).

1.1.4.3.7 Characterisation of alternative splicing for medical application

Abnormal RNA splicing is observed in a number of cancers, suggesting that modification of splicing may restore the phenotype and therefore offer a gene specific form of anticancer chemotherapy (125). Deregulation of RNA splicing can be induced in three ways:

- Mutations or polymorphisms within the gene (e.g. β-globin, and dystrophin)
- Changes induced because of alterations or mutations in the splicing machinery (e.g. SMN1)
- Changes in cell signalling pathways (e.g. FGFR2)

ß-thalassemia, which is caused by mutations in the ß-globin gene, is one of the most common human genetic diseases. The most frequent mutations are in introns 1 and 2 which result in abnormal splicing of the mRNA. Blocking the incorrect splice sites with antisense oligonucleotides has been demonstrated (126) and could provide a therapeutic approach to correct this disorder.

Duchenne muscular dystrophy (DMD) is caused by a frameshift in the dystrophin gene, which results in the insertion of a premature stop codon and the expression of a truncated inactive protein. Antisense oligonucleotides have been used to induce alternative splicing of dystrophin gene (127).These alternative splice events remove the exon that encodes the stop codon and return the transcript to the correct reading frame, thereby generating a near full length transcript that has partial activity and could reduce the severity of the condition.

Spinal muscular atrophy (SMA) is a disorder in which the patient suffers from progressive loss of spinal cord motor neurons which results in paralysis. This disorder occurs due to the deletion of the survivor of motor neuron gene (SMN1). SMN1 is involved in the assembly of ribonucleoprotein complexes (128).

Deletion of SMN prevents assembly of the U1 ribonucleoprotein complexes in the cytoplasm and results in global defects in pre-mRNA splicing.

Fibroblast growth factor receptor 2 (FGFR2). Activation of FGFR2 influences mitogenesis and differentiation. Two of the alternative RNA isoforms of FGFR2 (IIIb and IIIc) show differing affinities to the different fibroblast growth factors and therefore activation results in differential signalling. Changes in the FGFR2 IIIb and IIIc ratio correlate with prostate cancer progression and are associated with lung, skin and bone defects (129,130). Alternative splicing of FGFR2 is controlled via the polypyrimidine tract binding protein (PTB), which acts as an exon-silencing factor that is thought to have a role in the progression of cancer (129).

1.1.5 Further levels at which genes are regulated

a) RNA transport

A fully processed mRNA must leave the nucleus in order to be translated into protein (46).

b) Transcript Stability

RNA Longevity can also dictate when a gene is expressed. If two mRNA molecules are transcribed, one lasts for five minutes in the cytoplasm before being degraded, while the other one manages to linger for an hour before being degraded. If both are translated continually while they exist, more of the second polypeptide will be produced than the first, which is the principle behind regulation of RNA longevity. mRNAs from different genes have their approximate lifespan encoded in them; this serves to help regulate how much of each polypeptide is produced.. Eukaryotic mRNAs can vary greatly in their stability (46). The information for lifespan is found in the 3' untranslated region of the mRNA. For those more unstable RNAs this information signals for rapid degradation.

c) Translation Initiation

Since many mRNAs have multiple methionine codons, the ability of the ribosomes to recognise and initiate synthesis from the correct AUG codon can affect the expression of a gene product (46).

1.1.6 Recent developments in gene regulation

1.1.6.1 Small interfering RNA (siRNA) and micro RNA (miRNA)

RNA silencing is a well-studied mechanism of gene regulation that limits the transcript level by suppressing transcription (transcriptional gene silencing (131)). More recently a second method of gene silencing has been discovered. This mechanism activates a sequence-specific RNA degradation process which is called posttranscriptional gene silencing (PTGS) or RNA interference (132) (133).

During the past few years the discovery of two classes of short RNA molecules, siRNA and miRNA, 21-25 nucleotides in length involved in gene regulation have stunned scientists. Double-stranded RNA (dsRNA) directs the sequence-specific degradation of mRNA through a process known as RNA interference (RNAi). Only now are scientists beginning to explore the molecular processes responsible for RNAi and to appreciate the breadth of its function in biology (132).

RNAi is a two-step process. The first step involves degradation of dsRNA into small interfering RNAs (siRNAs) by an RNase III-like activity. In the second step, the siRNAs join an RNase complex, RISC (RNA-induced silencing complex), which acts on the corresponding mRNA and degrades it (Figure 1-12). Several key components such as Dicer, RNA-dependent RNA polymerase, helicases, and dsRNA endonucleases have been identified as being part of the RISC (134). Some of these components also control the development of many organisms by processing many noncoding RNAs, called micro-RNAs. At the moment it is believed that the function of miRNAs and siRNAs are interchangeable yet they are quite distinct complexes (135). Due to its specificity and efficiency, RNAi is being considered as an important tool not only for functional genomics, but also for gene-specific therapeutic activities that target the mRNAs of disease-related genes.

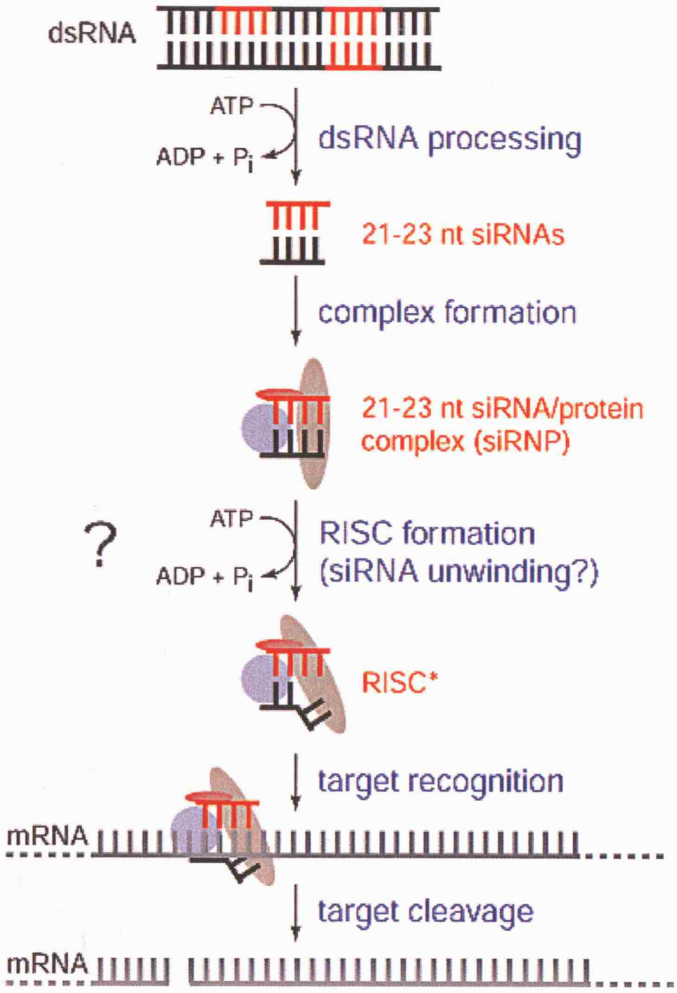

Figure 1-12 A model for the RNAi pathway. RNAi is a two-step process. The first step involves degradation of dsRNA into small interfering RNAs (siRNAs) by an RNase III-like activity. In the second step, the siRNAs join an RNase complex, RISC (RNA-induced silencing complex), which acts on the corresponding mRNA and degrades it. ATP-dependent cleavage of dsRNA yields siRNAs, which then guide a nuclease to the target mRNA, ensuring its destruction. The complex of proteins formed on the duplex siRNA is denoted 'siRNP' to distinguish it from the fully active 'RNA-induced silencing complex' (RISC*) capable of cleaving its RNA target. Figure taken from Zanmore 2001 (133).

1.1.7 Communication between chromosomes ("kissing" chromosomes)

Our understanding of the mechanism by which genes are regulated has evolved from the initial linear view of one gene, one promoter, one protein (Figure 1-13a), to the more complex view that proteins bound to DNA sequences hundreds of kilo-bases away can affect gene regulation (Figure1-13b). The most recent view is that genes on different chromosomes can affect the expression of one

another (Figure 1-13d). There is now evidence that the spatial location of a gene within a cell nucleus can determine whether a gene is expressed or not (136).

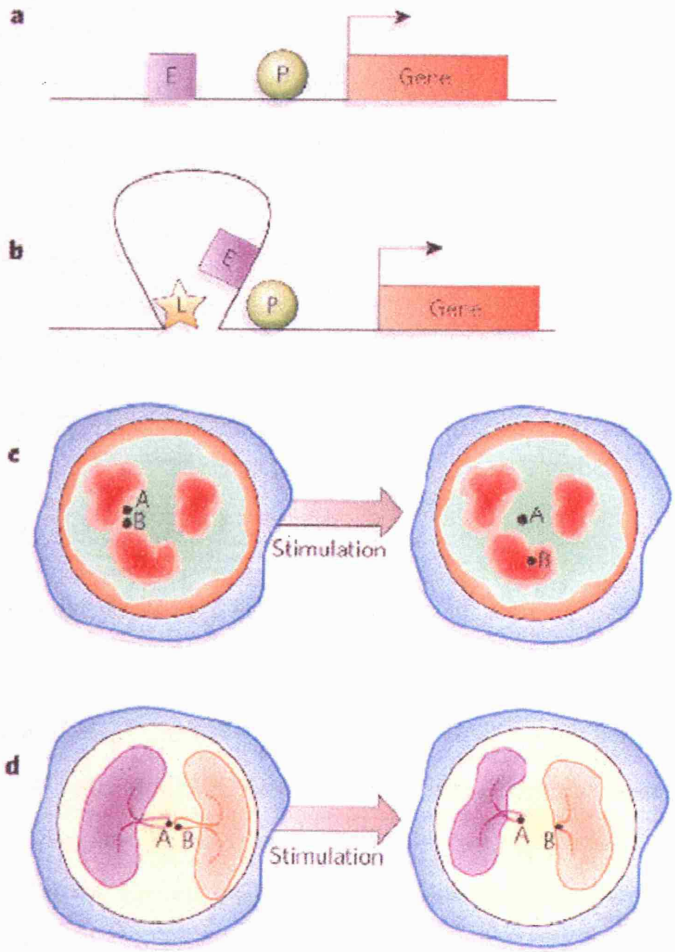

Figure 1-13 The three dimensions of gene regulation. (a) Linear view of gene regulation. The promoter (P) near the start of the gene provides the minimal information needed for gene expression. Regulatory proteins bind to enhancers or silencers (E) and supplement the function of the promoter, either by activating or repressing transcription of the gene (arrow). **(b) The 'looping scanning' model of gene regulation.** The locus control region (L) regulates several genes. Proteins bind here scan through large portions of DNA, looping out the intervening region, until they find the relevant gene. **(c) Gene regulation in 3D.** Genes from different chromosomes (A and B) are in close proximity until a developmental signal stimulates the cells, when the genes split apart. One moves to a region that represses gene expression (Heterochromatin, red) and the other relocates to an area with any active genes (euchromatin, green) **(d) Genes from different chromosomes in contact.** Genes from different chromosomes might come into contact when the chromatin containing them loops out from their chromosome 'territory'. Figure taken from Kioussis, Nature 2005 (136).

Post-stimulation, T helper cells (T_H) develop into either T_H1 or T_H2 cells which produce different effector molecules. T_H1 cells produce Ifnγ, whose gene is located on chromosome 10. T_H2 cells produce Il3, Il4 and Il5 genes located in a

cluster on chromosome 11. Biochemical and imaging analysis showed that the two regions (the Ifnγ gene and the Il3, 4, 5 gene cluster) on the two different chromosomes (10 and 11) are in close proximity in the nucleus of naïve T_H cells. After stimulation, that induces either a T_H1 or T_H2 state, they move apart (136). Spilianakis *et al.* believes that in the naive T_H cell all the genes are poised for expression. When the cell receives a stimulus, for example for T_H1 cellular differentiation, the gene to be activated, Ifnγ, is permitted to be transcribed. At the same time the T_H2 specific genes are moved to a more repressed region of the nucleus therefore preventing their expression (137). The exact mechanism behind the moving chromosomes is still not understood and probably will not be until 4 dimensional technologies are used to video the chromosomal interactions as they occur.

1.1.8 Non-polyadenylated RNAs

Across the evolutionary spectrum, transcription of DNA into RNA and the subsequent maturation of the RNA into a functional form are steps at which gene expression may be regulated. 98-99% of the human genome is non-coding DNA. This includes the introns found between protein-coding exons in genes, as well as the regions of DNA up and downstream of the genes themselves. It has long been known that genes are transcribed and processed before being transported from the nucleus to the cytosol. One vital step in RNA processing is the polyadenylation of the nascent transcript. Recent findings suggest that non-coding regions are also transcribed into polyadenylated (poly A+), stable RNAs that are transported into the cytosol during development. TUF (transcripts of unknown function) is an unofficial name given to these transcripts. The discovery of TUFs, the ever-increasing number of transcripts lacking 3' polyadenylation and the revelation that many genes are transcribed as poly A+ but then under certain conditions are processed to remove the 3' poly A sequences, indicate that DNA previously considered to be "junk" may in-fact play a crucial role in the regulation of some genes (138).

1.2 Protein Kinases

This class of enzymes, comprising several hundred different genes, is one of the largest protein families in the human genome. Indeed, about 1-2% of human genes encode for protein kinases which regulate most of the functions in metabolism, growth and differentiation of various cell types. They mediate the response of eukaryotic cells to external stimuli by phosphorylation of hydroxyamino acids.

Protein kinases are divided in two groups following their specificity: One group phosphorylates proteins on the amino acids serine and/or threonine; the other one phosphorylates proteins on tyrosine. However, some protein kinases are able to phosphorylate serine, threonine as well as tyrosine residues (10).

Protein phosphorylation is the most general regulatory mechanism in eukaryotic cells. Moreover, dysregulated protein kinases play key roles in major human diseases and therefore are a major group of drug targets in the pharmaceutical industry (139).

1.2.1 Tyrosine Kinases

Protein tyrosine kinases (PTKs) are enzymes that catalyse the transfer of the γ-phosphate of ATP to a tyrosine residue of protein substrates. They can be subdivided into two large families namely receptor tyrosine kinases (RTKs) (140) and non-receptor tyrosine kinases (NRTKs) (141).

1.2.1.1 Receptor Tyrosine Kinases

RTKs span the plasma membrane and contain an extracellular portion, which binds the ligand, and an intracellular portion, which possesses catalytic activity and contains the regulatory sequences. The RTK family includes the insulin receptor and the receptors for many growth factors such as epidermal growth factor (EGF), platelet-derived growth factor (PDGF), fibroblast growth factor (FGF) and nerve growth factor (140).

The binding of ligand to the extracellular portion of RTKs leads to dimerization of the monomeric receptors resulting in autophosphorylation of specific tyrosine residues in the cytoplasmic portion (Figure 1-14) (140).

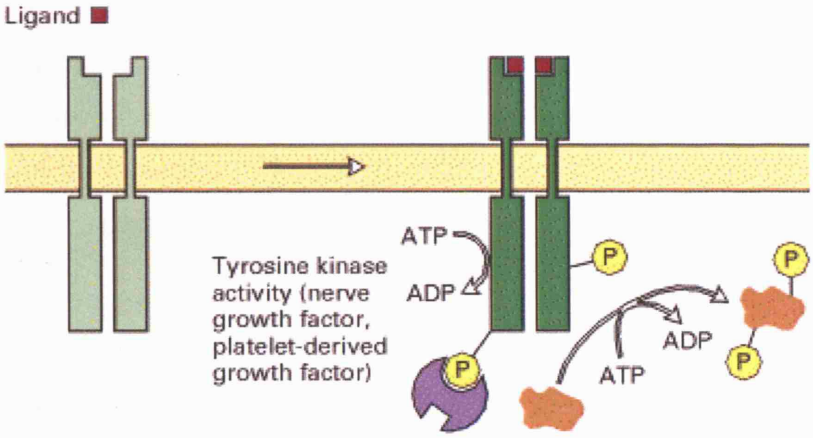

Figure 1-14 The receptors for many growth factors have intrinsic protein-tyrosine kinase activity. RTKs span the plasma membrane and contain an extracellular portion, which binds the ligand, and an intracellular portion, which possesses catalytic activity and contains the regulatory sequences. The RTK family includes the insulin receptor and the receptors for many growth factors such as epidermal growth factor (EGF), platelet-derived growth factor (PDGF), fibroblast growth factor (FGF) and nerve growth factor. Ligand binding to most such receptor tyrosine kinase causes formation of an activated homodimer, which phosphorylates several residues in its own cytosolic domain as well as certain substrate proteins. Tyrosine autophosphorylation either stimulates the natural catalytic (kinase) activity of the receptor or generates recruitment sites for downstream signalling proteins containing phosphotyrosine-recognition domains. Figure taken from Lodish *et al.* (10,142).

Tyrosine autophosphorylation either stimulates the natural catalytic (kinase) activity of the receptor or generates recruitment sites for downstream signalling proteins containing phosphotyrosine-recognition domains, such as the Src homology 2 (SH2) domain (143) (described in more detail in section 1.3).

Signalling proteins, which bind to the intracellular domain of receptor tyrosine kinases in a phosphotyrosine-dependent manner, include RasGAP, PI3K, phospholipase Cγ, phosphotyrosine phosphatase SHP and adaptor proteins such as Shc, Grb2 and Crk.

1.2.1.2 Nonreceptor tyrosine kinases

In contrast to RTKs, NRTKs are located in the cytoplasm, nucleus or anchored to the inner leaflet of the plasma membrane. They are grouped into eight families: SRC, JAK, ABL, FAK, FPS, CSK, SYK and BTK.

NRTKs contain no extracellular or transmembrane portion but possess modular domains that are responsible for subcellular targeting and regulation of catalytic activity. Because of the key roles PTKs play in cellular signalling processes, their catalytic activity is tightly controlled in normal cells by protein-tyrosine phosphatases (141).

1.3 Phospho-inositide 3-kinases (PI3Ks)

The PI3K family comprises three classes of enzymes (Class I, Class II and Class III) that catalyse the transfer of the γ-phosphate group of ATP to the D-3 position of the inositol ring of inositol phospholipids. These lipids lead to the recruitment of certain signalling proteins to the plasma membrane via their pleckstrin homology domain, described further in section 1.6.1 (144).

Both the class I and class II PI3Ks are an integral part of receptor-mediated signalling pathways prevalent in animal cells. In contrast, plants and yeast contain only class III PI3Ks. PI3Ks are at the route of a number of signalling cascades that control proliferation; cell death, migration, growth and a number of other biological processes. In recent years they have been implicated in the formation of tumours, metastasis (145), chronic inflammation, cardiovascular disease (146) and allergy (147). This knowledge of their involvement in disease makes the PI3-kinase family an increasingly attractive target for drug development.

The catalytic subunit of all members share a homologous region that contains a catalytic core linked to the PIK, PI kinase homology domain, and a C2 domain (Figure 1-15). PI3K can be grouped into 3 classes based on their structure, selective *in vitro* substrate specificity and regulation (144):

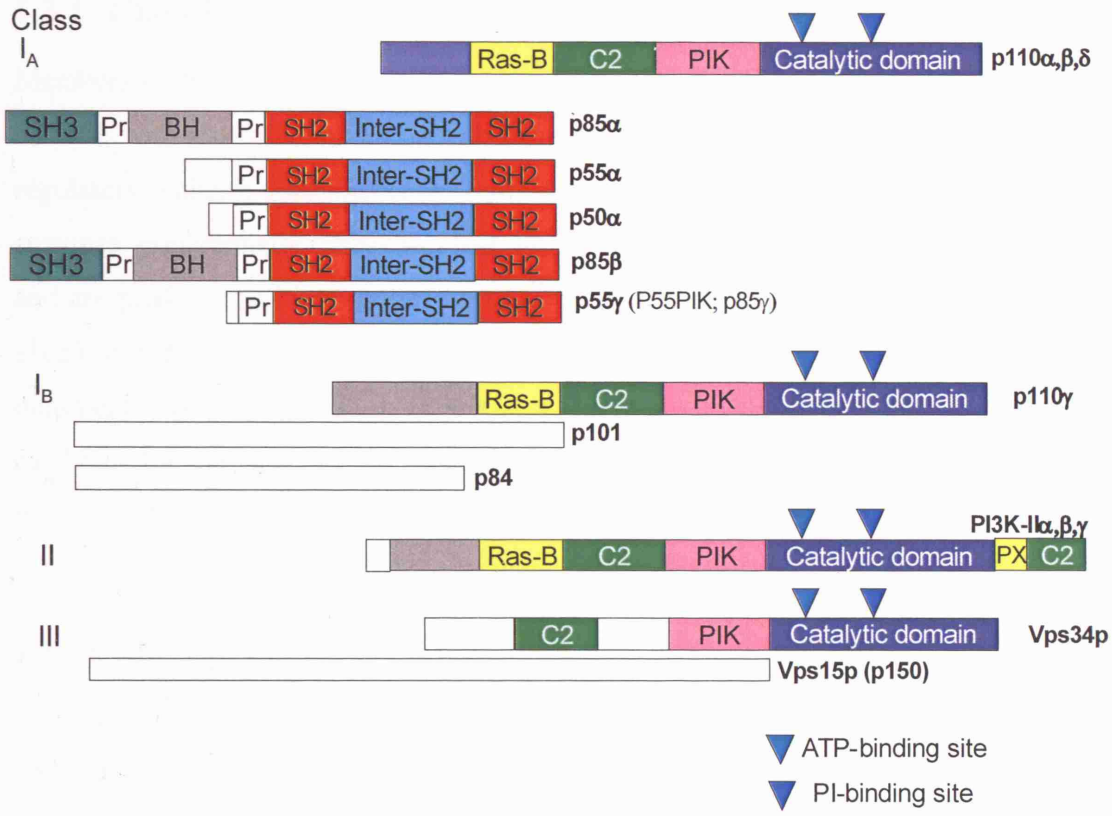

Figure 1-15 Modular structures of the three classes of PI3-kinase. The catalytic subunit for each class is shown on top of the corresponding regulatory subunit(s). Ras-B is the Ras binding domain, PIK is the PI kinase homology domain, SH domains are Src homology domains and Pr are proline rich domains. All three classes of PI3Ks contain a PIK a C2 and a catalytic domain. **Class IA PI3Ks** are heterodimeric proteins consisting of a p110 catalytic subunit and a regulatory subunit that has SH2 domains. Mammals express 3 Class IA p110 catalytic subunits (p110α, p110β and p110δ), encoded by 3 distinct genes. These p110s bind to one of 5 class IA regulatory subunits (p85α p55α, p50α, p85β and p55γ that link the catalytic subunits to tyrosine kinase signalling pathways. All three catalytic subunits can make functional complexes with all of the splice variants of the adaptor subunit. The adaptor proteins contain two Src homology (SH2) domains which bind to the phosphorylated tyrosine residues on the activated receptor. The inner SH2 domain mediates tight binding of the p85 regulatory subunit to the catalytic subunit. The only **class IB PI3K** to be identified to date is p110γ. It is similar in structure and function to the class IA p110 proteins except for the lack of the p85 regulatory subunit-binding domain. The 110 kDa catalytic subunit binds a 101 kDa regulatory protein or a recently characterised 84 kDa subunit. As a heterodimer, p110γ is activated by Gβγ subunits of heterotrimeric G proteins. There are three members of the **class II PI3K** family, PI3K-C2α, β and γ, all encoded by three separate genes. They are larger proteins (170 – 210 kDa) whose catalytic domain share 45-50% similarity with class I PI3Ks. The N-terminal region of these proteins has a C2 domain but other than that show no homology to any other known protein. The **class III PI3K** enzyme has a yeast homologue, Vps34p, which was identified as a yeast vesicular-protein-sorting protein, a protein essential for accurate transport of newly synthesized proteins from the Golgi to the vacuole. (Figure adapted from Vanhaesebroeck *et al.* (1)).

1.3.1 Class I PI3K

Members of the Class I PI3K family are heterodimeric proteins with a catalytic subunit, with a theoretical size of approximately 120 kDa (known as p110) and a regulatory subunit (schematically represented in Figure 1-15). *In vivo* these enzymes preferentially phosphorylate PtdIns(4,5)P_2 generating PtdIns(3,4,5)P_3 and are predominantly cytosolic. These enzymes require translocation, usually by a cell surface receptor, for full activation (1). They are sensitive to the PI3K inhibitors wortmannin and LY294002 which bind to the active site in their catalytic subunit and inhibit their kinase activity (148). Depending on the regulatory protein class I PI3Ks can be divided into two groups:

1.3.1.1 Class IA PI3K

Class IA PI3Ks are heterodimeric proteins consisting of a p110 catalytic subunit and a regulatory subunit that has SH2 domains. Mammals express 3 Class IA p110 catalytic subunits (p110α, p110β and p110δ), encoded by 3 distinct genes (*PIK3CA*, *PIK3CB* and *PIK3CD*, respectively). These p110s bind to one of 5 class IA regulatory subunits (p85α p55α and p50α, encoded by the *PIK3R1* gene and p85β and p55γ (encoded by the *PIK3R2* and *PI3KR3* genes, respectively) that link the catalytic subunits to tyrosine kinase signalling pathways. All three catalytic subunits can make functional complexes with all of the splice variants of the adaptor subunits (149). The adaptor proteins contain two Src homology (SH2) domains (Figure 1-15) which bind to phosphorylated tyrosine residues generated by activated receptor tyrosine kinases and receptor-associated adaptor proteins. The SH2 domains belonging to the these adaptor subunits preferentially bind to tyrosine residues in a specific Y(P)xxM motif, where x is any amino acid. This binding process allows translocation of the cytosolic PI3K to the cell membrane where it comes in close proximity to its lipid substrates and Ras (Figure 1-16) (148). The class IA adaptors are very diverse in their N-terminal regions which may contribute to differences in their protein binding partners, their cellular distribution and the level of activation of the catalytic subunits in response to different stimuli (148).

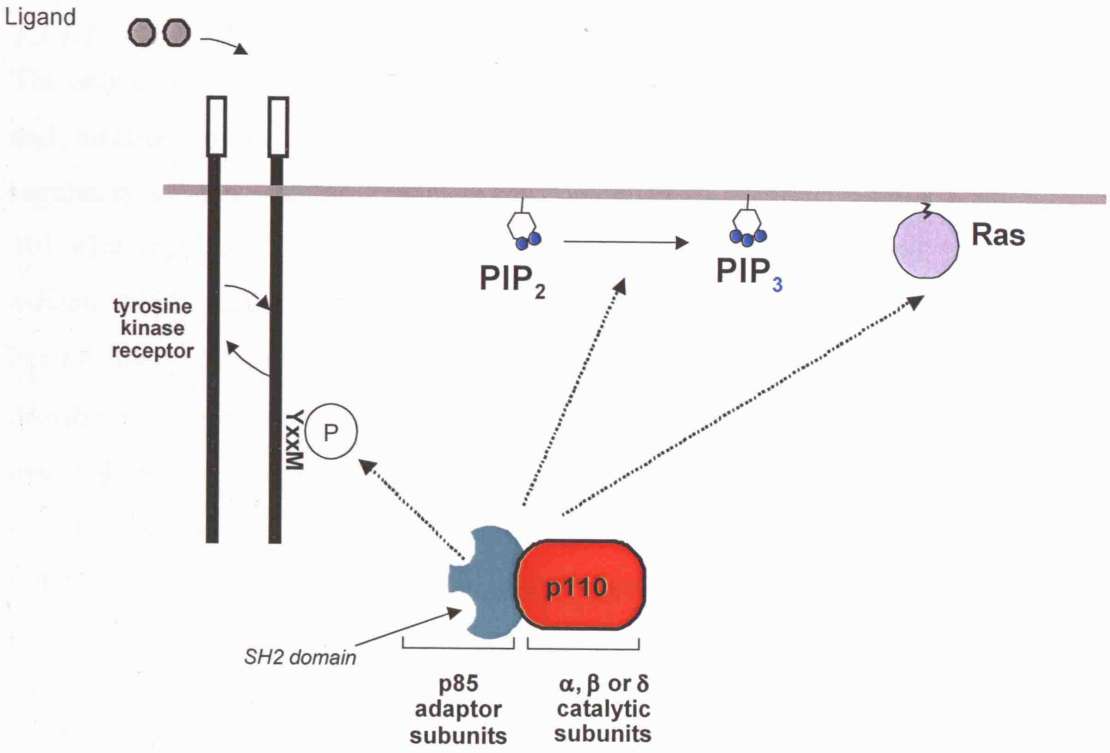

Figure 1-16 Recruitment/activation of Class IA PI3K. A receptor with intrinsic tyrosine kinase activity is shown to dimerize upon binding of its cognate ligand and to transphosphorylate, creating recognition/docking sites for the SH2 domains of the adaptor subunits of class IA PI3Ks. Apart from phosphorylating themselves receptor tyrosine kinases can also create phosphotyrosine docking sites on other molecules such as adaptor proteins (not shown). In some cases, the receptor or associated proteins are phosphorylated by cytosolic tyrosine kinases (not shown). Class I PI3K also interact with Ras but this interaction is not necessarily simultaneous with the phosphotyrosine interaction. (Figure adapted from Vanhaesebroeck *et al.* (148)).

Class IA PI3Ks regulate many cellular processes downstream of tyrosine kinase and Ras, including cell metabolism, growth, proliferation, differentiation, motility, and survival. Evidence is available suggesting that these isoforms have non-redundant functions such as in cell proliferation (150), survival, actin cytoskeletal reorganization (151) and migration (9).

p110δ is predominantly expressed in leukocytes, unlike the more broadly expressed p110α and p110β (8). p110δ expression has also been documented in breast cells and melanocytes. It has also been demonstrated that p110δ plays an important role in the EGF-driven migration of breast cancer cells (7), the development and differentiation of B cells, the differentiation and survival of effector and memory T cells and is an important mediator of antigen receptor signalling in B and T cells (152).

1.3.1.2 Class IB PI3K

The only class IB PI3K to be identified to date is p110γ. It is similar in structure and function to the class IA p110 proteins except for the lack of the p85 regulatory subunit binding domain (153). The 110 kDa catalytic subunit binds a 101 kDa regulatory protein (Figure 1-15) or a recently characterised 84 kDa subunit (154). As a heterodimer, p110γ is activated by Gβγ subunits of heterotrimeric G proteins. p110γ, like p110δ shows a restricted tissue distribution, being abundant only in white blood cells (1). P110γ plays an essential role in a number of physiological responses, including neutrophil chemotaxis, mast cell degranulation, and cardiac function. Research done by Laffargue *et al.* (155) revealed a cooperation between G protein coupled receptors and FcεRI signalling, suggesting that mast cell responses are intensified by signals through GPCR requiring p110γ.

1.3.2 Class II PI3Ks

There are three members of the class II PI3K family, PI3K-C2α, β and γ, all encoded by three separate genes (148). They are larger proteins (170 – 210 kDa) whose catalytic domain share 45-50% similarity with class I PI3Ks. The N-terminal region of these proteins have a C2 domain but other than that show no homology to any other known protein (5). *In vitro*, class II PI3K have a preference for PtdIns as a substrate but have been observed to phosphorylate PtdIns(4)P and PtdIns(4,5)P_2. A number of activating ligands have been identified such as insulin, epidermal growth factor, platelet derived growth factor and integrins. The molecular mechanism by which class II enzymes are activated is still unclear (148).

1.3.3 Class III PI3K

The class III PI 3-kinase enzyme has a yeast homologue, Vps34p, which was identified as a yeast vesicular-protein-sorting protein, a protein essential for accurate transport of newly synthesized proteins from the Golgi to the vacuole (153). A serine/threonine kinase, Vps15, has been identified in the recruitment of Vps34 to the membrane and enhancing its lipid kinase activity. A human VPS15

has also been cloned and identified as an adaptor subunit of the class III PI3Ks. PtdIns is the only class III substrate identified so far (5).

1.4 Phosphoinositides

Phosphoinositides (PI) are a family of phospholipids originating from phosphatidylinositol, PtdIns. The family include all phosphorylated forms of this lipid, PtdIns (PI), PtdIns(4)P, PtdIns(4,5)P_2, PtdIns(5)P, PtdIns(3)P, PtdIns(3,5)P_2, PtdIns(3,4)P_2 and PtdIns(3,4,5)P_3. The levels of phosphoinositides in mammalian cells vary greatly, PtdIns being the most abundant and PtdIns(3,4,5)P_3 and PtdIns(3,4)P_2, being the least abundant. Phosphoinositides are dynamic and vital members of the cell's repertoire of signalling molecules. They are involved in regulating diverse processes, such as cytoskeletal organization, membrane trafficking, and ion transport (156). Their intracellular level is strictly regulated by specific kinases, phosphatases and phospholipases (157).

In eukaryotic cells, the multiple phosphorylated forms and the specific lipid kinases involved in their synthesis (155) 3-kinases, PtdIns 4-kinases and PtdInsP kinases, are located in various intracellular compartments, including the plasma membrane, the cytoskeleton, or the nucleus.

1.4.1 Structure

Phosphoinositides are acidic phospholipids that consist of a hydrophobic tail, composed of the two fatty acid chains, making up the phosphatidic acid backbone, linked via a phosphate group to a hydrophilic polar inositol head group (Figure 1-17) (1,11).

The free hydroxyl groups at positions 3,4 and 5 on the inositol head group have been shown to be phosphorylated in different combinations, (discussed further in section 1.5). Phosphorylation of the hydroxyl groups at positions 2 and 6 has not been documented to date (1).

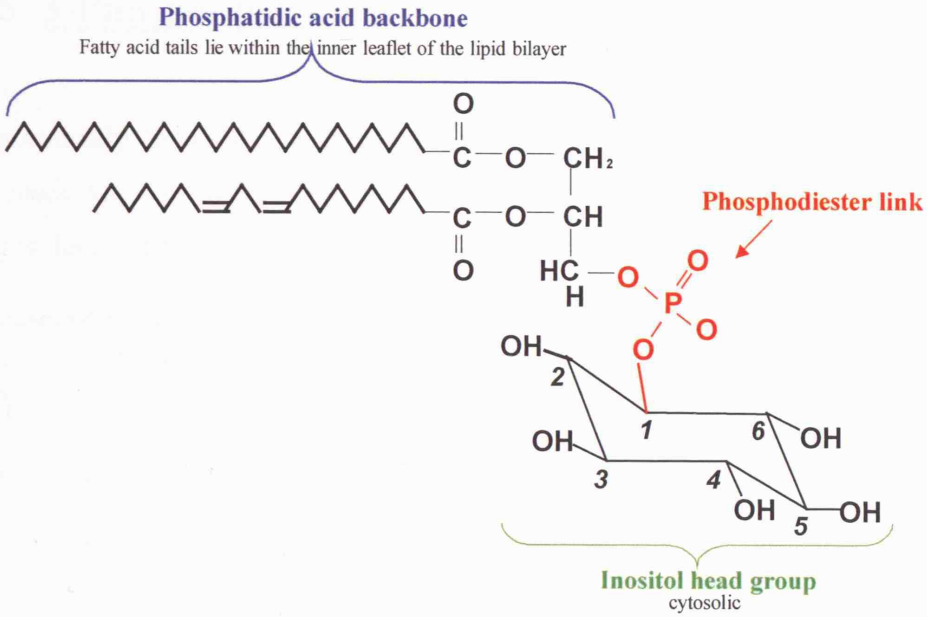

Figure 1-17 **Chemical structure of PtdIns**. Phosphoinositides are acidic phospholipids that consist of a hydrophobic tail, composed of the two fatty acid chains, making up the phosphatidic acid backbone, linked via a phosphate group to a hydrophilic polar inositol head group. The free hydroxyl groups at positions 3,4 and 5 on the inositol head group have been shown to be phosphorylated in different combinations. Phosphorylation of the hydroxyl groups at positions 2 and 6 has not been documented to date. Figure adapted from Vanhaesebroeck *et al.* (1).

1.4.2 Function

Biological membranes consist largely of bilayers of phospholipids to which proteins or protein complexes are attached or in which they are embedded. Besides their structural role as the building blocks of a membrane, some phospholipids can act as regulators of proteins or other phospholipids in the membrane (11). The inositol-containing phospholipids, or phosphoinositides, are an example of phospholipids with structural and regulatory functions. Phosphoinositides are involved in the regulation of diverse cellular processes, including proliferation, survival, cytoskeletal organization, vesicle trafficking, glucose transport, and platelet function (5).

As phosphoinositides are polyanionic, they can be very effective in non-specific electrostatic interactions with proteins. However, they are especially effective in binding specific lipid binding domains (1) (discussed further in section 1.6.1).

1.5 3-Phosphoinositides

Theoretically any of the combinations of 3-phosphorylated phosphoinositides can be made via the kinase activity of any member of the Class I PI3Ks, in reality this is does not happen.

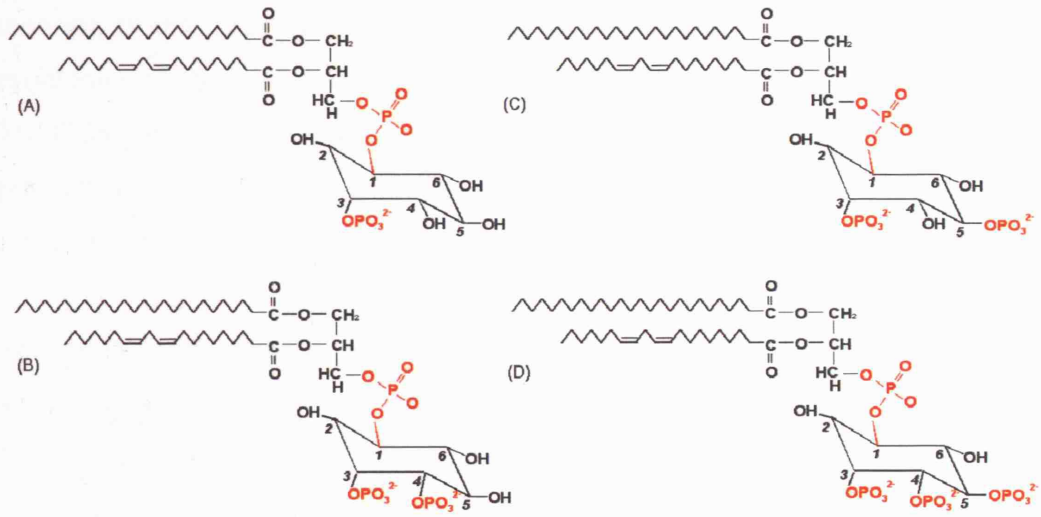

Figure 1-18 3-phosphorylated phosphoinositides (A) PtdIns(3)P (B) PtdIns(3,4)P_2 (C) PtdIns(3,5)P_2 (D) PtdIns(3,4,5)P_3. PI3Ks can be divided into three main classes on the basis of their structure, the mechanism by which they are activated and their *in vitro* lipid substrate specificity. Class I PI3K preferentially phosphorylate PI(4,5)P_2 to make PI(3,4,5)P_3 (**D**). Class II and III preferentially phosphorylate PI to give PI(3)P (**A**). PI(3,5)P_2 (**C**) is not a direct product of a PI3K, this lipid is synthesized by phosphorylation of the hydroxyl group on the 5 position of the inositol head group of PtdIns(3)P by a 5-kinase. There is evidence that PI(3,4)P_2 (**B**), is derived by 4-phosphorylation of PI(3)P. Figure adapted from Vanhaesebroeck *et al.* (148)

PI3Ks can be divided into three main classes on the basis of their structure, the mechanism by which they are activated and their *in vitro* lipid substrate specificity (149), as discussed in section 1.3. Class I PI3K preferentially phosphorylate PI(4,5)P_2 to make PI(3,4,5)P_3 (Figure 1-18d), and Class II and III preferentially phosphorylate PI to give PI(3)P (Figure 1-18a). PI(3,5)P_2 (Figure 1-18c) is not a direct product of a PI3K, this lipid is synthesized by phosphorylation of the hydroxyl group on the 5 position of the inositol head group of PtdIns(3)P by a 5-kinase. There is evidence that PI(3,4)P_2 (Figure 1-18b), is derived by 4-phosphorylation of PI(3)P, but a second scenario suggests the action of a 5-phosphoinositide phosphatase on PI(3,4,5)P_2 (148,158). The only characterised PI(3,4,5)P_3 3-phosphatase to date is the tumour-suppressor

gene PTEN which negatively regulates, basal and hormone-stimulated levels of $PI(3,4,5)P_3$ (158).

1.6 Roles of PI3K lipid products

The 3-phosphoinositide lipids, generated by PI3Ks, are important second messengers that regulate a broad variety of cellular responses such as growth, proliferation, survival, differentiation, intracellular traffic and cell migration (1-7). PI3K activity is critical in a wide variety of normal and pathological physiological responses, including immune regulation and inflammation, metabolic control and cancer.

The most recent findings include the involvement of PI3Ks in growth, DNA synthesis, apoptosis, the actin cytoskeleton and directional cell migration (Figure 1-19). Detailed experiments have demonstrated that the over-expression of the class IA PI3K (Dp110) in *Drosophila* results in increases in the growth of cells. On the other hand reduction in the PI3K activity levels has a negative effect on growth (159). Similar effects have been demonstrated in mice hearts when increasing/decreasing p110α. An increase in growth must be accompanied by an increase in protein synthesis one signalling pathway downstream of PI3K is that of p70-S6K, known to promote protein synthesis (1).

Activation of the platelet derived growth factor receptor (PDGFR) results in the stimulation of DNA synthesis in certain cell lines. Administration of a PI3K inhibitor abrogates this effect. Numerous proteins are required for cell cycle progression. PI3K activation has been demonstrated to influence the accumulation of certain proteins involved in cell cycle progression via increasing mRNA stability and mRNA translation as well as preventing protein degradation (1).

PI3K blocks apoptosis by regulating Akt/PKB. Akt/PKB is the cellular homologue of v-Akt, which causes cell transformation. It is thought that Akt/PKB mediates some cellular responses, including protection from apoptosis. PI3Ks are thought to down regulate cell apoptosis by stimulating Akt/PKB activity which can block apoptosis.

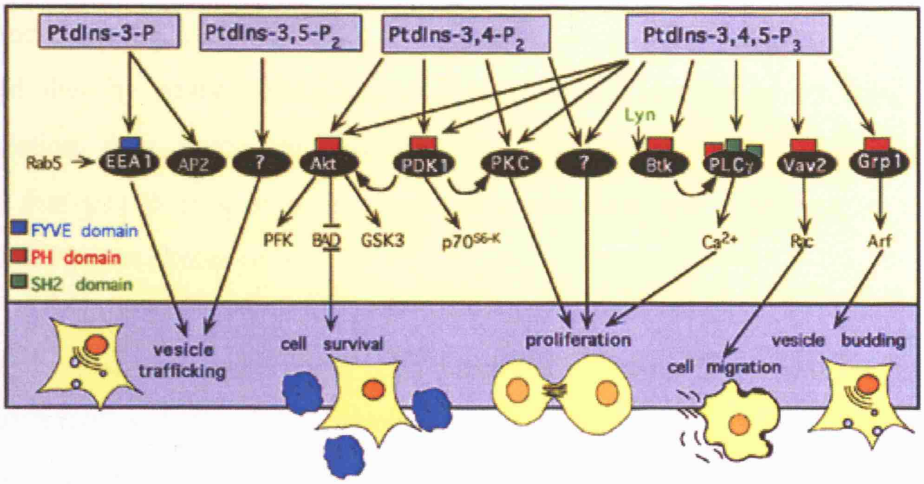

Figure 1-19 Signalling through PI3K lipid products and their targets. The lipid products of PI3Ks are indicated at the top of the figure, and the cellular processes affected by these lipids are indicated at the bottom. The black ovals are the direct targets of each lipid, and the small boxes indicate the protein domains that directly bind to them; the red boxes correspond to PH domain containing proteins, the blue boxes correspond to FYVE domain containing proteins and the green boxes correspond to SH2 domain containing proteins. The major lipid product produced via the activation of Class IA PI3Ks, PtdIns-3,4,5-P_3, plays a key role in the activation of a large number of downstream signalling pathways including cell survival, proliferation, cell migration and vesicle budding. Figure taken from Rameh *et al.* (6).

PKB activity is dependent on growth factors, which activate PI3K (160). Activated Akt/PKB phosphorylates BAD which, when phosphorylated, releases Bcl-2 allowing it to inhibit apoptosis. Caspase 9, when phosphorylated by Akt/PKB, is inhibited from initiating apoptosis. Akt/PKB regulation of NF-κB has also been implicated in the regulation of apoptosis. NF-κB is a transcription factor that protects cells from apoptosis by inducing the expression of anti-apoptotic genes (1).

It is not clear how PI3Ks affect the actin cytoskeleton but it is believed to be in a regulatory manner via guanosine nucleotide exchange factors (GEFs) and GTPase activating proteins (GAPs), which are all PH domain-containing proteins. Rac, a GTPase partially dependent on PI3K activation, is a major actin-remodeling factor. Other PI3K downstream effectors are ADP ribosylation factors (Arfs), members of a GTPase family involved in the regulation of intracellular membrane traffic and reorganization of the actin cytoskeleton (1).

The non-redundant role of certain members of the PI3K family in directional movement has been demonstrated by a number of groups. Sadhu *et al.* (161) showed that by administering a p110δ selective inhibitor to neutrophils, cell polarization, thus directional movement was inhibited. Sawyer *et al.* (7) also found that p110δ protein is important in chemotaxis and migration of EGF stimulated breast cancer cells.

Different PI3Ks substrates have been identified as being involved in different cellular activities. PtdIns(3)P is involved in trafficking of proteins within the cell. PtdIns(3,4,5)P_3 has been implicated in signal transduction that leads to adhesion by interacting with the PH and C2 domains of proteins, actin rearrangement, and cell growth and cell survival. PtdIns(3,4)P_2 has been implicated in cells growth and cell survival by affecting PKC, PDK and PKB (162).

1.6.1 PI3K lipid targets

The two main structurally distinct targets of 3-phosphoinositides are those containing a Pleckstrin Homology (PH) domain and those containing a FYVE domain, named after the first four proteins shown to contain it (1). Other protein modules that can bind lipids are the PHD finger domain, the PX domain and the ENTH domain.

Pleckstrin homology (PH) domains are 100-120 amino acid proteins units, found in over 250 proteins in the human genome, and are known to bind specific phosphoinositol phospholipids, most bind weakly and non-specifically (163). Out of the 250 known PH domain containing proteins twenty bind strongly to PI(3,4,5)P_3 and/or PI(3,4)P_2, all of which contain a specific sequence motif (164). These PH-domain containing proteins that recognise 3-phosphorylated phosphoinositides, drive membrane recruitment in response to PI3K activation (163).

The most well characterised PH domain mediated signalling pathway involves PKB/Akt and PDK1. PKB is cytosolic in unstimulated cells, where there are very low levels of PtdIns (3,4)P_2 or PtdIns (3,4,5)P_3. Upon activation of certain PI3Ks

there is an increase in these lipids resulting in the translocation of PKB to the plasma membrane, where it becomes activated. Although membrane recruitment and phosphoinositide binding have been reported to increase PKB activity directly, its phosphorylation by phosphoinositide-dependent protein kinase 1 (PDK1) has been shown to be required for full activation *in vivo*. Active PKB then detaches from the membrane and, using an unknown mechanism, translocates to the nucleus where it exerts some of its downstream effects. PKB plays a major role in promoting cell survival via a number of methods, discussed earlier in section 1.6. As well as cell survival PKB plays an important role in insulin signal transduction. PI3K inhibition prevents nearly all-physiological responses of mammalian cells to insulin. Other targets of PKB, downstream of insulin receptor signalling are GSK3, phosphodiesterase-3B (PDE-3B), mammalian target of rapamycin (mTOR), and insulin receptor substrate-1 (IRS-1) (165).

A second important signalling pathway that involves PH domain mediated recruitment to PI3K products is seen via the B-cell antigen receptor (BCR), and involves Btk. Mutations in the PH domain of Btk are known to cause impaired B-cell development (164). Membrane recruitment of Btk, via its PH domain, brings it into close proximity of Lyn, a Src-family kinase, present at the membrane. Lyn phosphorylates Btk, which then undergoes autophosphorylation to become completely active. Other target proteins that bind PtdIns(3,4,5)P_3 via their PH domains include Vav, a Rac GEF (guanine nucleotide exchange factor), Sos, a Ras/Rac GEF, GAP1, a RAS GAP (GTPase activating protein) and PLCγ_1, involved in the hydrolysis of PI(4,5)P_2 (166).

Proteins containing FYVE domains are more selective in their binding specificity. They bind solely to PI(3)P and are much fewer in number (1).

1.7 PI3K and disease

1.7.1 PI3Ks and cancer

The PTEN tumour suppressor gene encodes a 3-phosphoinositide phosphatase that counteracts PI3K activity. Inactivating mutations or deletions of the PTEN gene result in hyper-activation of the PI3K/Akt signaling pathway and are very common in human malignancies, including breast, ovarian, colon and pancreatic cancer (Table 1-1). Prior studies in different tumor models have shown that, under conditions of PTEN deficiency, the PI3K/Akt signaling pathway becomes a permenantly active, proliferative and survival pathway, and that pharmacological inhibition of this pathway results in tumor growth inhibition (146,167).

Cancer Type	Type of alteration
Glioblastoma	PTEN mutation
Ovarian	Allelic imbalance and mutation of PTEN Elevated Akt 1 kinase activity Akt 2 amplification and over expression PI3K p110α amplification PI3K p85α mutation
Breast	Elevated Akt 1 kinase activity Akt 2 amplification and over expression Ribosomal protein S6K amplification and over expression Loss of heterozygosity at PTEN locus PI3K and Akt 2 over activation
Endometrial	PTEN mutation PTEN silencing
Hepatocellular carcinoma	PTEN mutation
Melanoma	PTEN mutation PTEN silencing
Digestive tract	Aberrant PTEN transcripts PI3K p85α mutation
Lung	PTEN inactivation
Renal cell carcinoma	PTEN mutation
Thyroid	PTEN mutation Akt over expression and over activation
Lymphoid	PTEN mutation

Table 1-1Evidence of PI3K signalling deregulation in human malignancies. The PTEN tumour suppressor gene encodes a 3-phosphoinositide phosphatase that counteracts PI3K activity. Inactivating mutations or deletions of the PTEN gene result in hyper-activation of the PI3K/Akt signaling pathway and are very common in human malignancies, including breast, ovarian, colon and pancreatic cancer. Prior studies in different tumor models have shown that, under conditions of PTEN deficiency, the PI3K/Akt signaling pathway becomes a permenantly active, proliferative and survival pathway, and that pharmacological inhibition of this pathway results in tumor growth inhibition. Table taken from Vivanco *et al.* (167).

1.7.2 PI3K and diabetes

Class IA PI3K activity is necessary for many of the actions of insulin including the stimulation of glucose transport, activation of glycogen synthase, stimulation of protein synthesis and the inhibitory effects on lipolysis (Figure 1-20). Activation of PI3K function is impaired in the insulin resistant state that precedes type-2 diabetes (168).

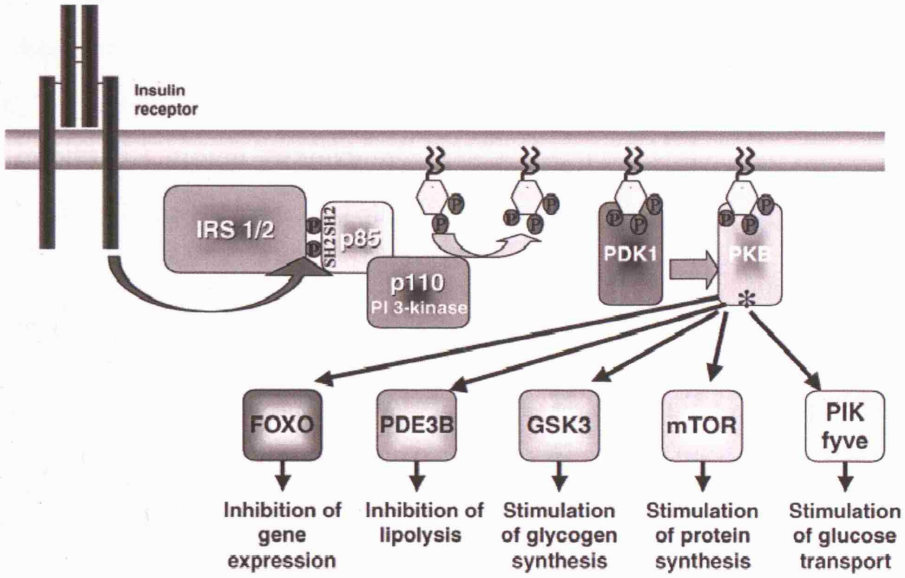

Figure 1-20 Insulin signalling pathways via class IA PI 3-kinase. Stimulation of cells with insulin results in a 10-100 fold increase in the concentration of PtdIns (3,4,5)P3 at the plasma membrane. PKB is activated following its recruitment to the plasma membrane of cells in a PI3K dependent manner. FOXO controls insulin-dependent cellular differentiation. Foxo1 plays an important role in adipocyte development and physiology. Insufficient Foxo1 has been shown to decrease adipocyte size and increase expression of genes that promote lipid metabolism. Foxo1 phosphorylation is tightly regulated. Expression of a phosphorylation-defective Foxo1 prevents adipocyte differentiation by altering expression patterns of genes involved in cell cycle control and adipogenesis. The antilipolytic effect of insulin is mediated through the activation of the cGMP-inhibitable phosphodiesterase PDE3B. This enzyme becomes activated by phosphorylation on serine 302 through an insulin-regulated serine kinase which has yet to be identified. PDE3B, a protein of around 135 kD, is associated with the plasma membrane and is dependent on PI3-kinase for its activation. However, it is not known whether PI3-kinase itself or other downstream kinases are involved in the phosphorylation and activation of PDE3B. Insulin induces glycogen synthesis by stimulating the phosphorylation and subsequent inactivation of glycogen synthasae kinase 3 (GSK3). Inactivation of GSK3 permits the continued action of glycogen synthase, the enzyme that catalyses the final step of glycogen synthesis. GSK3 phosphorylation is mediated by PKB. Protein kinases p70S6K and mTOR (mammalian target of rapamycin), which are downstream of PDK1 and PKB, have been identified as insulin-induced kinase that inhibit IRS-1. The PI3K/PKB pathway phosphorylates IRS-1 and inhibits its ability to bind PI3K and activate PKB; this effect is sensitive to rapamycin, an inhibitor of mTOR. Figure taken from Shepherd *et al.* (168).

Insulin increases glucose transport in skeletal muscle by inducing the translocation of GLUT4 (the major insulin-regulated glucose transporter) from intracellular vesicles to the plasma membrane. Expression of the GLUT4 gene is normal in skeletal muscle from patients with Type II diabetes. It has been demonstrated that it is defective insulin action that is the cause of impaired trafficking and function of GLUT4 (169).

Binding of the regulatory subunit of PI3K, at Src homology 2 (SH2), to the insulin receptor substrates (IRS) results in activation of PI3K. This step is necessary for insulin action on glucose transport. PI3K activation is responsible, at least in part, for insulin stimulation of GLUT4 translocation from intracellular vesicles to the plasma membrane (168,169).

1.7.3 PI3K and autoimmune diseases

The specific role of individual PI3Ks is yet to be determined but p110γ and p110δ being predominantly expressed in leukocytes (8) indicates that they may play unique roles in immune signalling.

Mutant mice lacking p110α and p110β are embryonic lethal (1). The viability of the p110δ null mouse (152,170) enabled scientists to investigate the role of p110δ *in vivo*. Mice with catalytically inactive p110δ fail to mount a normal immune response and the animals showed substantial defects in the activities of the B and T lymphocyte cells, two crucial components for initiating an immune response. These mice also developed inflammatory bowel disease (152).

A study into the allergic response produced by mast cells identified p110δ as the main mediator of the release of inflammatory substances from these cells. Upon an encounter with an allergen, in complex with immunoglobulin IgE, an aggregate is formed with the Fcε receptor on mast cells activating a cascade of biological reactions that ultimately lead to the release of proteins involved in the allergic response, such as histamines and cytokines, known as degranulation (10). Ali et al demonstrated that FcεRI activation was substantially reduced in

p110δ null cells. In the bone marrow macrophages of $p110\delta^{D901A}$ (loss of function) mutant mice, degranulation was reduced by 45-55% compared to that of wild type mice. They showed defects in the secretion of inflammatory cytokines as well as the failure of SCF, usually a critical component for mast cell degranulation, to initiate degranulation, thus demonstrating the primary role for p110δ in IgE dependent degranulation. Thereby showing that inactivation of p110δ protects mice against anaphylactic allergic response (147).

Mice lacking the p85α subunit also showed impaired B cell development and activation but normal T cell activation (171). A lethal lymphoproliferative disease was diagnosed in mice with a T-cell restricted deletion of PTEN, the phosphatase involved in the negative regulation of class I PI3Ks (172).

1.8 Therapeutic potential of PI3Ks

The involvement of PI3Ks in the regulation of many cellular processes has become very clear over recent years. The involvement of their key negative regulator, PTEN, in numerous cancers has brought them to the forefront of investigation as a target for drug development. A number of PI3K pathway inhibitors are in clinical trials (Table 1-2).

Although there is strong evidence suggesting that inhibiting class I PI3Ks may be beneficial in the treatment of cancer and autoimmune diseases, their role in numerous major signalling pathway, and their, still unresolved, non-redundant functions, means that as it stands, broad spectrum inhibitors would result levels of toxicity that may result in adverse toxicity in humans.

The most exciting molecule in the PI3K field at the moment is p110δ. It is clear that p110δ plays a unique role in antigen receptor signalling (152) suggesting that a specific inhibitor of p110δ could be a useful therapeutic intervention to B and T cell mediated autoimmunity and allergy as well as possible B and T cell transformation. Identification of the regulatory mechanism of p110δ gene expression may play a critical role in the development of p110δ selective inhibitors.

Drug name	Target	Company	Indication	Clinical trial phase
CCI-779	mTOR	Wyeth-Ayerst	Advanced breast cancer and renal cell carcinoma	Phase III
RAD001	mTOR	Novartis	• Prosate cancer • Advanced non-small-cell lung cancer • Gleevec-resistant gastrointestinal stromal tumours	Phase I/II
AP23573	mTOR	ARIAD pharmaceuticals	Advanced sarcoma and relapsed or refractory haematological malignancies	Phase II
Perifosine	PKB/Akt	Keryx pharmaceuticals	• Refractory non-small-cell lung cancer • Recurrent prostate cancer	PhaseII

Table 1-2 Phosphoinositide 3-kinase pathway inhibitors in clinical trials. A number of PI3K pathway inhibitors are in clinical trials. **CCI-779** is an ester of the immunosuppressive agent rapamycin that causes cell-cycle arrest at G1 via inhibition of key signaling pathways resulting in inhibition of RNA translation. Antitumor activity has been demonstrated using cell lines and animal models of malignant glioma. **RAD001** is a novel protein kinase mTOR-inhibitor. Activation of the PI-3K/AKT/mTOR pathway has been frequently associated with prostate cancer as well as advanced non-small-cell lung cancer. Studies are ongoing to determine dose and dosing schedule effects of RAD001 on tolerability and molecular modifications of mTOR pathway signaling in patients with newly-diagnosed prostate cancer requiring prostatectomy. **AP23573**, a novel analog of rapamycin, inhibits mTOR signaling in tumors, leading to cell cycle arrest, tumor cell shrinkage and inhibition of angiogenesis. Mutations that affect tumor suppressor genes, such as PTEN, are proposed to enhance sensitivity to mTOR inhibition. PTEN regulates cell growth by inhibition of the PI3K pathway, which promotes protein translation and cell cycle progression. **Perifosine** has displayed significant antiproliferative activity *in vitro* and *in vivo* in several human tumor model systems and has recently entered phase II clinical trials. It causes dose-dependent inhibition of protein kinase B/Akt phosphorylation and thus activation at concentrations causing growth inhibition of PC-3 prostate carcinoma cells. Only the myristoylated form of Akt (MYR-Akt), which bypasses the requirement for pleckstrin homology (PH) domain-mediated membrane recruitment, abrogated perifosine-mediated decrease of Akt phosphorylation and cell growth inhibition by perifosine. Table taken from Stokoe *et al.* (173).

1.9 Aim

Thousands of manuscripts on the subject of PI3Ks have been published in the last 5-10 years. Despite the importance of this signalling system, very little is known about the regulation of PI3K gene expression under normal and diseased conditions. The only detailed study performed was on the p110γ catalytic subunit class IB PI3K (174). Detailed analysis of the p110γ gene identified tissue-specific promoter activity within a 675 bp sequence upstream of the ATG translation start site (174).

The detailed tissue distribution of p110α and p110β in mammals is unknown, but all evidence indicates that these isoforms show a wider distribution than p110δ and p110γ, which are predominantly found in white blood cells. The mechanism of gene expression of the p110 subunits and their regulatory subunits is unknown. Evidence for post-translational regulation of protein stability has been reported for the p110α class IA PI3K catalytic subunit. p110α appears to be thermally unstable, and is conformationally stabilized and its activity inhibited by binding to the p85α regulatory subunit (175). Evidence for an increase in the copy number of PIK3CA has been reported in certain cancers, such as squarmous cell carcinomas (176), cervical (177), ovarian (178) and lung cancer (179). A report has also shown an increase in copy number of PIK3CD, aswell as PIK3CA, in glioblastomas (180). An increase in the expression of p85α PI3K in human isolated adipocytes upon activation of PPARγ by Rosiglitazone has recently been reported (181). This activation and increase in expression of p85α had no affect on the insulin receptor, insulin receptor substrate (IRS)-1 or the p110α and β catalytic subunits of PI3K. Several reports have indicated induction of expression of the p55α subunit, without simultaneous induction of the p110 subunits (182-184). This indicates that expression of the catalytic and regulatory subunits can be independent of each other.

Treatment of U937 leukaemia cells with all-trans retinoic acid has been reported to increase p110γ gene transcription, without effect on the levels of p110β, p110δ, p85 and p101 proteins (185).

There is strong evidence that suggests alterations in PI3K function could play a pivotal role in the pathogenesis of diseases such as diabetes and cancer. Normalisation of Class IA-PI3K activity (i.e. increasing it in diabetes or inhibiting it in cancer) may offer new paths for therapeutic intervention.

This project aims to identify the regulatory mechanism involved in the transcription of the Class IA p110δ catalytic subunit of the PI3K family, in the hope that it will help towards the development of treatments for immunological diseases, such as asthma, or towards the treatment of diabetes or even cancer.

2 Materials and Methods

2.1 General Materials

General chemicals were obtained from Sigma Aldrich Company Ltd., Poole, Dorset and B.D.H. Ltd., Poole, Dorset.

All cell culture media and reagents were obtained from GibcoBRL, Life Technologies Ltd., Paisley, Scotland.

All molecular biology reagents such as restriction enzymes, Klenow Fragment, T4 DNA ligase, Calf Intestine Alkaline Phosphatase, T4 Polynucleotide Kinase, and Mung Bean Nuclease were purchased from MBI Fermentas, Hanover, MD, USA or from New England Biolabs, Beverly Massachusetts, USA

All bacterial culture media was obtained from Becton Dickinson, Maryland, USA.

Cell lines were brought up from frozen stocks in the laboratories' liquid nitrogen storage tanks or purchased from ATCC, Teddington, UK.

In general, buffers and solutions for DNA, protein and cell work were made up using deionised water, and where possible, were autoclaved or filtered through 0.2 μm filters. Where ultra pure water was required, the water was also sterilized by autoclaving.

2.2 Cell culture

2.2.1 Materials

1) Complete culture media:

DMEM: DMEM (Dulbecco's modified Eagle's medium) (GibcoBRL) supplemented with penicillin (100 units/ml medium), streptomycin (100 μg/ml medium) and 10% (v/v) FCS (foetal calf serum)

RPMI: RPMI-1640 (Roswell Park Memorial Institute) (GibcoBRL) supplemented with penicillin (100 units/ml medium), streptomycin (100 μg/ml medium), 10% (v/v) FCS (foetal calf serum) and 0.05 mM 2-mercaptoethanol.

2) ACK lysis buffer: 8.29 g NH_4Cl (0.15 M); 1 g $KHCO_3$ (1.0 mM); 37.2 mg Na_2EDTA (0.1 mM); Adjust to pH 7.2-7.4 with 1 M HCl and filter sterilise.

Cell line	Cell type	Organism
A20	Pre B cell leukaemia	mouse
Baf 3	Pre B cell	mouse
B16-BL6	melanoma	mouse
B16-F10	melanoma	mouse
CCRF-CEM	T cell leukaemia	human
CT26	epithelial	mouse
EAHy 926	HUVEC derived endothelial	human
EL4	T cell leukaemia	mouse
EMT 6	breast malignant	mouse
HELA	cervical carcinoma	human
HEK 293	human embryonic kidney	human
Jurkat	Tcell lymphoma	human
LLC	epithelial	mouse
L929	fibroblast	mouse
MB431	breast malignant	human
MDA-MB 231	breast cancer	human
MDA-MB 361	breast cancer	human
MDA-MB 468	breast cancer	human
MEF	embryonic fibroblast	mouse
melan-b	melanoma	mouse
NIH3T3	fibroblast	mouse
NMuMG	breast-benign	mouse
Swiss 3T3	fibroblast	mouse
THP 1	Acute monocytic leukaemia	human

Table 2-1 A list of all cell lines used in this thesis

Cell culture media, PBS and trypsin were all kept at 37°C to warm prior to handling of the cells. The filtered air cabinet was switched on 10 min before use and cleaned thoroughly with 70% ethanol. All cell culture work was carried out with sterile glass and plastic tissue culture grade pipettes and flasks.

2.2.2 Culturing cells from frozen stocks

Cells were removed from liquid nitrogen and thawed at 37°C. Cells from the cryovial were resuspended in 5 ml of complete medium. To remove the DMSO from the solution, in which the cells were stored, the cells were centrifuged at 2,000 rpm for 5 min. The cell pellet was then resuspended in 10 ml of complete medium and transferred to a 25 cm^2 sterile tissue culture flask. The cells were kept at 37°C in a 5% CO_2 incubator and allowed to grow.

2.2.3 Maintenance of adherent cells

Growth of the cells was checked daily to prevent them becoming 100% confluent. Once 70-80% confluent, the cells were washed once with 1 x PBS before being trypsinised with 1 ml of trypsin at 37°C. Once the cells had detached from the substratum they were resuspended in 10 ml complete medium. 1 ml of the newly resuspended cells was transferred to a new 75 cm^2 tissue culture flask and topped up with 20 ml of fresh medium. The flask was then returned to the incubator.

2.2.4 Maintenance of suspension cells

Growth of the cells was checked daily to prevent them becoming too dense. Once they had reached optimum density the cells were centrifuged at 1,000 rpm for 5 min and the cell pellet was washed once in PBS then re-centrifuged. The washed pellet was resuspended in 10 ml complete medium. 1 ml of the newly resuspended cells was transferred to a new 75 cm^2 tissue culture flask and topped up with 20 ml of fresh media. The flask was then returned to the incubator and allowed to continue growing.

2.2.5 Preparation and maintenance of mouse spleen and thymus cells

For some experiments, we used cells and tissues directly isolated from mice. The mouse was dipped in ethanol before an incision was made to remove the spleen and thymus. The organ was then placed on a filter in a 60 ml tissue-culture dish containing 5 ml RPMI + 1% FCS. The organ was cut into pieces and pressed through the screen with a bulb from a 10 ml syringe. The medium with the cells was decanted into a 50 ml tissue-culture tube. The filter and the dish were washed with 5 ml of medium, which was added to the 50 ml tissue culture tube. The tube was then filled with ACK solution to lyse contaminating red blood cells, and allowed to incubate at room temperature for 5 min before being centrifuged at 1300 rpm (500 x g) for 7 min. The cell pellet was resuspended in 5 ml of complete RPMI.

2.2.6 Cryopreservation of cells

75-80% confluent cells were trypsinised then centrifuged (adherent cells), or just centrifuged (suspension cells) at 1500 rpm for 5 min. The pellet was resuspended in complete medium (depending on the number of cells) plus 1/10 volume of DMSO. 1 ml of the cells was aliquoted into labeled cryovials (NUNC). The aliquots were placed in a freezing container, put at -80°C overnight and then transferred to liquid nitrogen.

2.3 Restriction Digests, Electrophoretic Separation and Recovery of Nucleic Acids

2.3.1 Materials

Size markers called 'DNA ladders' were obtained from Fermentas Life Sciences

The sizes of these markers are as follows:

MassRuler™ DNA Ladder, High Range, ready-to-use:

10000, 8000, 6000, 5000, 4000, 3000, 2500, 2000, 1500 bp

MassRuler™ DNA Ladder, Low Range, ready-to-use:

1031, 900, 800, 700, 600, 500, 400, 300, 200, 100, 80 bp

1 x TAE buffer: 40 mM Tris.HCl pH 7.8, 20 mM sodium acetate, 1 mM EDTA

2.3.2 PCR

Solutions used for PCR were kept exclusively for this purpose and not stored near DNA or PCR products in order to reduce the risk of contamination. Preparations were carried out in clean conditions using filtered tips (Alpha Laboratories) for pipetting the solutions. In order to avoid excess handling of solutions, master mixes of the commonly used components were made up. Volumes for one reaction, as described in this section, were scaled up in accordance with the number of reactions being run. Pipetting error was taken into account by making master mix for one extra reaction. For example, enough master mix for 5 reactions was made if running 4 PCR reactions. The master mix was aliquoted into sample tubes.

The 10 x dNTP solution contained all four desoxynucleotides each at a concentration of 2 mM (MBI fermentas). Primers were resuspended to 100 pmole/µl with dH_2O. Annealing temperatures for the primer pairs varied according to the melting temperature (Tm), calculated using the formula 4(G+C) + 2(A+T). PCR was carried out in a PTC-200 Peltier Thermal Cycler – MJ research. This cycler has a heated lid which meant that the samples did not require a layer of mineral oil to prevent evaporation.

a) PCR on genomic DNA

x µl DNA (0.5-2 µg)
5 µl 10x PCR buffer
5 µl 2 mM dNTP
4 µl 25 mM $MgCl_2$
5 µl 5' primer @ 5 pmol/µl
5 µl 3' primer @ 5 pmol/µl
1 µl Taq DNA polymerase (5 U/µl)
25-x µl dH_2O
---
total volume of 50 µl

PCR cycles were as follows:

94°C for 3 min	x 1 cycle
94°C for 1 min 55-65°C for 1 min 72°C for 1 min	x 30-40 cycles
72°C for 4 min	x 1 cycle

b) Reverse transcriptase PCR (RT-PCR)

First Strand cDNA synthesis was carried out using the Superscript II Reverse Transcriptase system (Life Technologies). 5 µg of total RNA was incubated for 10 min at 65°C with 1 µl 500 µM oligo dT primer (MWG Biotech) and 10 mM dNTP mix (MBI Fermentas) in a total volume of 12 µl and then chilled on ice. After adding 4 µl of First Strand Buffer, 1 µl of DTT (100 mM stock), 2 µl of dNTP mix (10 mM stock) and 1 µl of Superscript reverse transcriptase (200 units/µl), the reaction was incubated for 50 min at 42°C and finally, for 5 min at 95°C. The original RNA template strand was removed by using 10 units of RNAseH per 5 µg RNA reaction. One tenth of the cDNA (2 µl) was used for subsequent PCR, with 39 cycles of amplification (paragraph 2.3.2), using exon-specific forward primers and a common reverse primer (Table 2-2). All RT-PCR reactions were repeated numerous times using RNA/cDNA prepared at different times. A forward primer to mouse exon 1 (Table 2-2) was designed as a positive control for RNA/cDNA quality. PCR products were separated on a 2% agarose gel and visualized by ethidium bromide staining. The

PCR products were cloned into pGEM®-T Easy (Promega) and sequenced using the T7 primer.

Primer name	Sequence 5'-3'	Annealing temp.	Application
M 1	cgtggttgttgacttcttgc	60	RT-PCR
M -1	caggcagaagtgggatgaa	58	RT-PCR
M -2a	gtagaaaggaaacaaagtgg	56	RT-PCR
M -2b	gaagacgaacagtgatgtag	58	RT-PCR
M -2c	gacccgccatgagcagttc	60	RT PCR
M RT rev	tgccaatgaggaggctgatc	56	RT-PCR
H -1	taaggagtcaggccagggcg	66	RT-PCR
H -2a	agtcgctccgagcggccgc	64	RT-PCR
H -2b	cgaggttgggagaggagtgtg	64	RT-PCR
H RT rev	cgggacacagggaagttcag	64	RT-PCR
adapt inner	cgcggatccgaacactgcgttt	70	5'RACE
adapt outer	gctgatggcgatgaatgaacactg	72	5'RACE
M-RACE inner	acttgaacttccccgtgtcccg	70	5'RACE
M-RACE outer	cagatcagctcctcattggcact	70	5'RACE
H-RACE inner	cgggacacagggaagttcaggt	70	5'RACE
H-RACE outer	gcttcttcacgcggtcgccc	68	5'RACE
pGEM RV3	ctagcaaaataggctgtccc	60	sequencing
pGEM GL2R	ggaagacgccaaaaacataaag	64	sequencing
-2b 2kb F	ggaggttacagatagacgcttcg	66	Promoter analysis
-2b 2kb R	ggcttacacttctacatcactg	64	Promoter analysis
-2b 150bp F	tagtgccatcagagaagaggg	64	Promoter analysis
-2b 150bp R	ggcttacacttctacatcactg (same as -2b 2kb R)	64	Promoter analysis
-2a 2kb F	cgaaggctggggattaaatagg	66	Promoter analysis
-2a 2kb R	gtccgcacactccacttccc	66	Promoter analysis
-2a 742bp F	catgggagtgggcttgcttc	64	Promoter analysis
-2a 742bp R	gtccgcacactccacttccc (same as -2a 2kb R)	66	Promoter analysis
-2a 479bp F	agattgtcccagtacacccag	64	Promoter analysis
-2a 479bp R	gtccgcacactccacttccc (same as -2a 2kb R)	66	Promoter analysis
-2a 256bp F	ttgaacagagctggggttcga	64	Promoter analysis
-2a 256bp R	gtccgcacactccacttccc (same as -2a 2kb R)	66	Promoter analysis
-1 2kb F	ggctaaaatgtgtctatctccgc	68	Promoter analysis
-1 2kb R	gggcttcatcccacttctgca	66	Promoter analysis
-1 989bp F	gtgaggagggttgcaagcaag	66	Promoter analysis
-1 989bp R	gggcttcatcccacttctgc (same as -1 2kb R)	66	Promoter analysis
-1 539bp F	tgtattctttctccgtgcgtgc	66	Promoter analysis
-1 539bp R	gggcttcatcccacttctgc (same as -1 2kb R)	66	Promoter analysis
1 2kb F	ttcacgggactgaaagacaaga	64	Promoter analysis
1 2kb R	ctgttgtgttacttctctgagg	64	Promoter analysis
1 446bp F	tggctttgctctcctcctgtt	64	Promoter analysis
1 446bp R	cctgttgtgttacttctctgaggtctg (same as 1 2kb R)	64	Promoter analysis
1 185bp F	gctatagggtggcttaaattgg	64	Promoter analysis
1 185bp R	cctgttgtgttacttctctgaggtctg (same as 1 2kb R)	64	Promoter analysis
1 91bp F	ctcatgtgtagcaagctttgca	64	Promoter analysis
1 91bp R	cctgttgtgttacttctctgaggtctg (same as 1 2kb R)	64	Promoter analysis
1 62bp F	agttcatcatgcctttgcctttc	66	Promoter analysis
162 bp R	cctgttgtgttacttctctgaggtctg (same as 1 2kb R)	64	Promoter analysis

Table 2-2 Primers used throughout this study for the indicated applications.

2.3.3 Restriction enzyme digestion

DNA and restriction endonuclease reaction components were pipetted into an Eppendorf tube, mixed briefly by vortexing then centrifuged to bring the contents to the bottom of the tube. Incubations were carried out at 37°C, except for *SmaI* single digests that were incubated at 30°C. Standard reaction conditions are listed below. In the case of the reaction containing more than one enzyme, the total volume of the enzyme did not exceed 10% of the reaction volume.

Plasmid DNA

x µl of DNA (0.5-4 µg)

2 µl of 10x restriction enzyme buffer

18-x µl dH_2O

Total volume 20 µl

Incubation 1-2 hr

2.3.3.1 Electrophoresis of DNA

Electrophoresis was carried out using two different gel systems:

Minigels were made dissolving 0.5 g of agarose in 50 ml 1 x TBE buffer (for a 1% gel). This was brought to the boil in a microwave and allowed to cool to 60°C. 2.5 µl of 10 mg/ml ethidium bromide was added and mixed thoroughly. This was then poured into a 10 cm x 6.5 cm gel casting frame. An 8-well comb was placed at the top end of the gel. The gel was allowed to set for 30 min before the comb was removed and the gel loaded into an electrophoresis tank filled with 1 x TBE. Samples were loaded into the wells and the electrophoresis was carried out at a constant voltage of 95 V for 1-2 h.

Midigels were made dissolving 3 g of agarose in 300 ml 1 x TBE (for a 1% gel). This was boiled and allowed to cool before 15 µl of 10 mg/ml ethidium bromide was added and poured into a 20 x 20 cm gel casting frame, sealed with autoclave tape. Two 16-well combs were placed in the frame, one at the top and one in the middle. The gel was allowed to set for 30 min before the tape was removed and the gel placed in an electrophoresis tank. The samples were loaded and the tank connected to 120 V for 1-2 h.

2.3.4 Extraction of DNA from agarose gels

The agarose gel was visualized under UV illumination and the required DNA fragment excised with a sterile scalpel blade and placed in a pre-weighed Eppendorf tube. The Eppendorf tube containing the gel slice was reweighed and the weight of the gel slice calculated. Five volumes of PCR binding buffer (PB, Qiagen) was added to the gel slice then heated to 50°C for 10-30 min until all of the gel had dissolved. The QIAquick DNA extraction protocol was used to purify the DNA from the agarose gel. This system relies on the DNA being absorbed to a silica membrane in the presence of high salt while the contaminants pass through the column. The agarose gel dissolved in the PB buffer was placed in a QIAquick spin columns and centrifuged for 1 min. The flow through solution was discarded and 750 µl of PE buffer was added to the column before it was centrifuged for another minute. The flow through solution was discarded once more and the column re-spun. It was then placed in a clean microcentrifuge tube and the DNA was eluted with 30 µl dH_2O.

2.4 Subcloning of DNA sequences

2.4.1 Bacteria, media, buffers and solutions

XL1-Blue strain: These cells are *endA* minus which improves the quality of miniprep DNA, and allow for blue/white colour screening.

Genotype: *rec*A1 *end*A1 *gyr*A96 *thi*-1 *hsd*R17 *sup*E44 *rel*A1 *lac* [F'*pro*AB *lacI*^{q}*ZΔ*M15 Tn*10* (Tet^{r})]

PGL3 Basic vector was purchased from Promega, Southampton. The pGL3 Luciferase Reporter Vector provides a basis for the insertion of DNA sequences and quantitative analysis of factors that act on these sequences to regulate gene expression. This vector was used to characterize the putative regulatory sequences of the p110δ promoter.

pGEM®-T Easy Vector System was purchased from Promega, Southampton. This vector was used for the cloning of PCR products generated by the Taq DNA polymerase. The single 3′-T overhangs at the insertion site improves the efficiency of ligation of a PCR product into the plasmids by preventing recircularisation of the vector and providing a compatible overhang for PCR products generated by Taq DNA polymerase. The vector also contains T7 and SP6 RNA polymerase promoters flanking a multiple cloning region within the α-peptide coding region of the enzyme β-galactosidase. Insertional inactivation of the α-peptide allows recombinant clones to be directly identified by colour screening on X-gal plates. The RNA promoters present in this plasmid were also used for transcription of antisense RNA probes

Plasmid DNA preparation kits were from QIAGEN Inc., Chatsworth, California, USA.

Ultraspec™-II RNA isolation system was purchased from Biotecx Laboratories, Inc., Houston, Texas, USA

Oligodeoxyribonucleotides were synthesized by QIAGEN Operon, Cologne, Germany.

L-Broth: 25 g LB powder in 1 litre of ddH_2O – autoclaved at 121°C for 15 min

LB-Agar plates: 25 g LB powder, 15 g Bacto Agar in 1 litre of ddH_2O – autoclaved at 121°C for 15 min

X-gal plates: IPTG (isopropyl b-D-thiogalactopyranoside): 40 µl of 100 mM stock in ddH20 (stored at -20°C), X-gal 40 µl of 20 mg/ml stock in DMSO (stored at -20°C) per LB-Agar plate

TfBI: 3 g 30 mM KAc; 50 ml 50 mM $MnCl_2$; 7.5 g 100 mM KCl; 1.5 g 10 mM $CaCl_2.2H_2O$; 150 ml 15% (v/v) glycerol. Make to 1 litre, filter sterilise and store at 4°C.

TfBII: 2.5 ml 10 mM NaMOPs (pH 7.0); 2.75 g 75 mM $CaCl_2$; 0.188 g 10 mM KCl and 37.5 ml 15% (v/v) glycerol. Make to 250 ml, filter sterilise and store at 4°C.

Phenol:Chloroform: an equal volume of water-saturated phenol and chloroform.

Chloroform: isoamyl-alcohol solution: 49 volumes of chloroform and 1 volume of isoamyl-alcohol.

3 M NaOAc: 408.1 g of sodium acetate anhydrous dissolved in 800 ml of dH_2O and brought to pH 5.2 with acetic acid, made up to 1 litre and sterilised by autoclaving.

70% ethanol: 70 ml of 100% ethanol with 30 ml dH_2O.

10 x TE buffer: 100 mM Tris.HCl pH 8.0; 10 mM EDTA.

SOC complete media: 0.5% yeast extract; 2% tryptone; 10 mM NaCl; 2.5 mM KCl; 10 mM $MgCl_2$; 10 mM $MgSO_4$; 20 mM glucose.

IPTG: 100 mM stock, made up in ddH_2O

X-gal: 20 mg/ml stock, made up in DMSO.

2.4.2 Production of competent cells

A stock of competent bacteria (XL1 Blue) was used to streak single colonies onto an LB plate. The plate was incubated at 37°C overnight. One colony was used to set up an overnight 20 ml culture at 37°C, shaking at 250 rpm. A one litre conical flask was used to set up a 400 ml LB media culture, inoculated with 1 ml of the 10 ml overnight culture of competent cells. The cells were left to grow until they had reached the log phase of growth (about 2-3 h). The cell density was measured by taking absorbance readings with a spectrophotometer at 600 nm. Cells within the log phase have an absorbance of between 0.5 and 0.7. Once the log phase had been reached the flask was placed into a large ice water bath and swirled gently, to cool the culture rapidly.

The culture was transferred to 4 x 250 ml Sorvall centrifuge bottles and centrifuged at 6,000 rpm for 20 min at 4°C. The supernatant was removed and the pellets were

resuspended in 25 ml ice cold TfBI (in each bottle – 100 ml in total) by pipetting up and down. The cells were re-pelleted by centrifugation at 6,000 rpm for a further 20 min. The supernatant was poured away and the pellets resuspended in 5 ml cold TfBII (per bottle – 20 ml in total). The newly made stock was aliquoted into 100 µl aliquots and stored at -80°C.

2.4.3 Subcloning of DNA fragments

a) Preparation of PCR products

The PCR products, created as in section 2.3.2a, were cloned into pGEM®-T Easy:

5 µl 2 x ligation buffer

1 µl pGEM®-T Easy vector

3 µl PCR product

1 µl T4 DNA ligase (3 Weiss units/µl)

The reaction was mixed and incubated at room temperature for 1 h. XL1-Blue competent *E coli* were transformed with the circularized plasmid as described in section 2.4.3fg). The pGEM®-T Easy vector contains the α-peptide coding region of the enzyme β-galactosidase. Insertional inactivation of the α-peptide allows recombinant clones to be directly identified by colour screening on X-gal indicator plates.

LB-Amp plates were coated with 40 µl (20 mg/ml) X-gal and 40 µl IPTG (100 mM) and left at 37°C. After 30 min incubation in SOC medium, the transformation mix was spread onto the LB-Amp + IPTG + X-gal plate and incubated overnight at 37°C.

White colonies from the plates were picked and grown overnight in 5 ml LB media (+ ampicillin) at 37°C, on a shaker at 250 rpm. The DNA was extracted using the method described in section 2.4.4.2.

One *EcoRI* restriction enzyme site either side of the cloning site was used (method described in section 2.3.3) to excise the PCR product. It was purified from the agarose using the QIAquick method (2.3.4) and then the 5' overhangs were filled in using Klenow DNA polymerase (2.4.3b).

b) Klenow treatment

In any situation where DNA fragments had 5' overhanging ends incompatible with the vector cloning site, these ends were filled in to create blunt-ended fragments using the protocol as follows: 2 µl of 10 x reaction buffer was added to 0.1-4 µg of DNA in 10-15 µl of dH_2O. 0.5 µl of 2 mM dNTP and 1-5 units of Klenow fragment were added to the sample and it was made up to 20 µl with dH_2O. The sample was incubated at 37°C for 10 min. followed by incubation at 70°C for 10 min. The PCR product, now blunt-ended was ligated into the PGL3 basic vector (2.4.3f).

c) Mung Bean Nuclease treatment

In any situation where DNA fragments had 3' overhanging ends incompatible with the vector cloning site, these ends were digested by Mung Bean Nuclease to create blunt-ended fragments using the protocol as follows: The DNA was resuspended in 1x Mung Bean Nuclease reaction buffer to a concentration of 0.1 µg/µl. 1 unit of Mung Bean Nuclease was added per µg of DNA. The sample was incubated at 30°C for 30 min followed by incubation at 70°C for 10 min. The DNA was recovered by ethanol precipitation. The PCR product, now blunt-ended was ligated into the PGL3 basic vector (2.4.3f).

d) Preparation of plasmid DNA

The PGL3 basic vector (Promega) was digested with *Sma I* (creates blunt DNA ends). Restriction enzyme digestions were carried out according to section 2.3.3. Once the plasmid had been digested, the sample was extracted using phenol:chloroform. An equal volume of phenol:chloroform was added to the sample and centrifuged in a microcentrifuge at maximum speed for 5 min. The aqueous layer was transferred to a clean Eppendof tube and mixed with an equal volume of chloroform:IAA and re-centrifuged at maximum speed for 5 min. The aqueous layer was precipitated with 1/10 volume of 3 M NaOAc and 2 volumes of 100% ethanol at –70°C for 1 h. The sample was then centrifuged at maximum speed for 15 min. The supernatant was discarded and the DNA pellet was washed with 70% ethanol and re-centrifuged at maximum speed for 5 min. The supernatant was discarded and the pellet left to dry at 37°C for about 5-10 min, until all of the ethanol had evaporated. The pellet was

resuspended in an appropriate volume of 1 x TE buffer. The DNA concentration was determined as described in section 2.4.4.5.

e) Phosphatase treatment

Plasmid DNA, used for sub-cloning, was subjected to phosphatase treatment to remove the phosphate group from the 5' end of the newly cut ends thus prevent the re-ligation of the blunt ends. The linearised plasmid was treated with 1 µl (1 unit/µl) of Calf Intestine Alkaline Phosphatase (CIAP) (MBI fermentas) and 1/10 volume of CIAP buffer at 37°C for 1 h. After the incubation the enzyme was heat inactivated at 65°C for 15 min.

f) Ligation reaction

The ligation reactions were set up between the de-phosphorylated PGL3 basic vector and either a PCR fragment or annealed ordered double stranded oligonucleotides. The concentration of the plasmid DNA used for ligation varied from 25-50 ng. Blunt-ended ligations were performed with a 2-4 fold molar excess of insert, depending on the availability of the DNA insert. The ligation reactions were kept to a minimum total volume. The reactions were set up with the appropriate controls including vector alone samples with and without ligase.

25-50 ng plasmid DNA
1/10 volume ligation buffer (MBI)
2-4 molar excess of insert
1/10 volume T4 ligase enzyme (5 U/µl)
appropriate volume of dH_2O.

The ligation reactions were incubated at room temperature overnight.

g) Transformation of ligated vectors

An appropriate number of aliquots of competent XL1-Blue cells (section 2.4.2) were thawed on ice. 0.5 µl of ligated DNA plasmid was added to each 100 µl aliquot and store on ice for 30 min. During this time one LB-amp plate per transformation was warmed at 37°C. The bacteria plus DNA was heat-shocked at 42°C for 45 sec and then returned to ice for 2 min. 900 µl of SOC media was added to each transformation and shaken at 250 rpm at 37°C for 1 h. 250 µl of each transformation was spread onto the LB-amp plates and incubated at 37°C overnight. The following morning the

colonies were picked and the DNA extracted using the method described in section 2.4.4.2.

h) Analysis of subcloning

10 µl of the purified DNA was digested (as described in section 2.3.3) using appropriate restriction enzymes to analyse the efficacy of the ligation and transformation. Successfully transformed samples were prepared for DNA sequencing (section 2.5) to confirm the sequence and the orientation of the insert.

2.4.4 Nucleic Acid Extraction

2.4.4.1 Large scale DNA extraction of plasmids

Qiagen MaxiPrep columns were used for the large scale DNA extraction of plasmids. A single colony was picked from a freshly streaked selective plate and used to inoculate a starter culture of 5 ml LB medium containing the appropriate selective antibiotic. The culture was incubated overnight at 37°C with shaking (250 rpm). 1 ml of the starter culture was used to inoculate 250ml of LB/Amp medium and grown at 37°C with shaking (250 rpm). The bacterial cells were harvested by centrifugation at 6,000 x g for 15 min at 4°C. The manufacturers' (Qiagen) protocol was then followed to extract the plasmid DNA.

DNA was precipitated by adding 10.5 ml isopropanol at room temperature, mixed and immediately centrifuged at 15,000 x g for 30 min at 4°C. The pelleted DNA was washed with 5 ml 70% ethanol, at room temperature, and centrifuged at 15,000 x g for 10 min at 4°C. The pellet was air-dried and dissolved in 300 µl of 1 x TE buffer. The DNA was stored at -20°C.

2.4.4.2 Small scale DNA extraction from plasmids

Qiagen MiniPrep columns were used for the large scale DNA extraction of plasmids. Bacterial cultures were established by transferring single, isolated colonies from LB-agar plates to L-Broth (2 ml) containing appropriate antibiotic (e.g. 100 µg/ml ampicillin) and growing overnight at 37°C in a shaking incubator.

The following morning the 2 ml cultures were transferred to micro-centrifuge tubes and spun for 15 sec. The manufacturers' protocol was then followed. The DNA was eluted in 30 µl of nuclease-free water by centrifuging at maximum speed for a further min. The DNA was stored at -20°C.

2.4.4.3 UltraspecII™ isolation of total RNA from cell lines

The UltraspecII™ RNA isolation system is based on the principle of a formulation of a 14 M solution of guanidine salts and urea (called Chaosolv), which act as denaturing agents and on the use of a specific RNA-binding resin to purify total RNA.

5-10 x 10^6 cells were lysed in 1 ml Ultraspec ™ RNA solution by repetitive pipetting. Following homogenization the samples were stored at 4°C to permit complete dissociation of nucleoprotein complexes. 0.2 ml of chloroform was added and the samples were shaken vigorously for 15 sec before being stored on ice for 5 min. The homogenate was centrifuged at 12,000 x g at 4°C for 15 min.

After centrifugation the homogenate forms two layers; a lower, organic phase and an upper, aqueous phase. The DNA and proteins reside in the organic phase while the RNA is in the aqueous phase. The aqueous phase was transferred to a fresh tube and mixed with 0.5 x its volume of isopropanol. 0.05 x sample volume of RNATrack™ resin was added and the tube vortexed for 30 sec.

The sample was spun for 30 sec before the supernatant was discarded. The pellet was washed twice with 1 ml 75% ethanol, vortexed and re-centrifuged for 30 sec each wash. The supernatant was discarded and the tube re-spun. Any traces of ethanol were removed using a pipette tip.

The RNA was resuspended in 1 volume (as that of the resin) of DEPC-treated water (Ambion) and vortexed for 30 sec before a 1 min centrifugation step.

2.4.4.4 Total RNA extraction from tissues using ToTALLY RNA™

This procedure is similar to the guanidinium thiocyanate/acid phenol:chloroform method but has been modified to include a phenol:chloroform extraction step to reduce or eliminate DNA, carbohydrate, heme, and other contaminants that can be a problem with some of the first generation guanidine-based methods.

2.4.4.4.1 Perfusion of mice to eliminate the circulating leukocytes

When using primary cells from mice, the animals were terminally anaesthetized using Pentobaribital Sodium (150 mg/kg). Their reflexes were tested to ensure they were fully anaesthetized and that they could feel no pain. After a bilateral incision of skin, peritoneum and ribcage, the right atrium was cut to let blood out and 20 ml of PBS was perfused through a 26G needle into the left ventricle. The organs were removed and placed directly into RNA *later*® buffer (Ambion).

RNA *later*® stabilizes and protects cellular RNA in intact, unfrozen tissue samples. Each organ was homogenized using a manual tissue homogenizer in 10 volumes of denaturing solution. To reduce the viscosity of the lysate it was passed through a 25-gauge syringe needle several times. The volume of the lysate was measured and recorded as the *Starting Volume*. 1 x *Starting Volume* of Phenol:Chloroform:IAA was added to the lysate. The sample was shaken vigorously for 1 min then stored on ice for a length of time which was dependent on the sample volume:

Sample volume (ml)	Incubation on ice (min)	Centrifugation (min)
1.5-2	5	5
10-15	10	15
> 15	15	15

Then the sample was centrifuged at 10,000 x g for the indicated time above at 4°C.

The upper, aqueous, layer was measured and transferred to a new Eppendorf tube. 1 x *Starting Volume* of Acid-Phenol:Chloroform was added to the sample and then it was shaken vigorously for 1 min before being stored on ice and centrifuged, following the guidelines in the table above.

The upper aqueous layer was measured and then transferred to a new RNase-free Eppendorf tube. An equal volume of isopropanol was added to the RNA sample and mixed well before being stored at -20°C for 30 min. The precipitation mixture was centrifuged at 12,000 x g for 15 min and the supernatant was carefully removed. The tube was re-spun and residual fluid removed.

300 µl of 70% ethanol, at room temperature, was added and the sample gently vortexed. The sample was then centrifuged at 3,000 x g for 10 min at 4°C. The ethanol supernatant was removed and the sample re-spun and the residual ethanol removed with a fine-tipped pipette. The pellet was resuspended in 100 µl of DEPC-treated water.

2.4.4.5 Quantification of DNA and RNA

The optical density (OD) of DNA or RNA obtained from extraction was determined at 260 nm and 280 nm using the NanoDrop ND-1000A Spectrophotometer. The NanoDrop works with 1 µl samples and eliminates the need for dilutions, cuvettes, and capillaries. It is capable of measuring extremely low amounts of nucleic acid and is more accurate than traditional spectrophotometers at high concentrations.

Knowing the optical density of DNA/RNA samples enables the concentration of the samples to be calculated, since:

OD_{260} of 1.0 = 50 µg/ml double stranded DNA

= 40 µg/ml single stranded DNA and RNA

The ratio of OD_{260} to OD_{280} of a pure DNA/RNA sample is 1.8. This allows the purity of the DNA/RNA to be determined at the same time. Contamination with protein would lower this ratio.

2.5 Automated DNA sequencing

2.5.1 Materials

DYEnamic™ ET* Terminator Cycle Sequencing Kits were purchased from Amersham Biosciences UK Ltd, Buckinghamshire.

Sequencing of double stranded DNA was initially performed 'in house' using an ABI PRISM 377 sequencer and the ABI PRISM dye terminator cycle sequencing ready reaction kit. Later it was sent to MWG (MWG Biotech, Germany) for commercial 'off site' sequencing. The primers used to sequence the cloned inserts in the PGL3 basic vector were RV3 and GL2R (Promega) described in Table 2-2. The reaction mix consisted of an optimized mixture of A-dye terminator; C-dye terminator; G-dye terminator and T-dye terminator, dATP, dCTP, dGTP, dTTP; Tris.HCl pH 9.0; $MgCl_2$; thermally stable pyrophosphatase and Ampitaq DNA polymerase.

The following solutions were mixed in an Eppendorf tube: 0.25-0.5 µg DNA template; 8.0 µl of termination-ready reaction mix; 3.2 pmol primer and the reaction was made up to 20 µl with dH_2O. The samples were loaded into the thermal cycler and subjected to the following thermal cycling:

96°C for 30 sec
50°C for 15 sec
60°C for 4 min
} 25 cycles
held at 4°C

The unincorporated dye terminators were removed by ethanol precipitation: 2 µl of 3 M sodium acetate pH 5.2 and 50 µl of 95% ethanol was added to the reaction. It was vortexed and placed on ice for 30 min and then centrifuged at maximum speed in a microcentrifuge for 10 min. The ethanol was carefully removed with a pipette and the pellet was washed with 250 µl of 70% ethanol and re-centrifuged at maximum speed for 3 min. The ethanol was removed and the pellet dried at 37°C for 5 min. The sample pellet was resuspended in 10 µl of formamide loading buffer.

2.6 Gene expression analysis

2.6.1 Materials

Dual-Luciferase® Reporter Assay System was purchased from Promega, Southampton.

BioRad Electroporator was purchased from Bio-Rad Laboratories Ltd., Hemel Hempstead, Hertfordshire.

4 mm BioRad cuvettes were purchased from Bio-Rad Laboratories Ltd., Hemel Hempstead, Hertfordshire.

2.6.2 Transfection

2.6.2.1 Transient transfection

a) Adherent cells

The day before transfection, 2.5 x 10^5 cells were seeded in a 60 mm tissue-culture dish in 5 ml appropriate culture medium and incubated overnight. On the day of transfection 5 μg of DNA (4 μg experimental firefly luciferase DNA and 1 μg *Renilla* luciferase DNA) was diluted with 150 μl cell growth medium without serum, proteins or antibiotics. The solution was mixed and centrifuged to remove drops from the top of the tube. 30 μl of SuperFect Transfection Reagent (Qiagen) was added and mixed by pipetting up and down 5 times. The sample was incubated for 10 min at room temperature to allow formation of the DNA-dendramer transfection complex.

While complex formation was taking place the growth medium was aspirated from the cells and they were washed once with 4 ml PBS. 1 ml cell growth medium, containing serum and antibiotics, was added to the transfection complex, mixed by pipetting up and down twice and immediately transferred to the cells in the 60 mm dish. The cells were incubated with the transfection complexes for 3 h under their normal growth conditions. The reaction medium was removed from the cells by gentle aspiration and the cells were washed once with 4 ml PBS. Fresh culture medium, containing serum and antibiotics was added and the cells were returned to their normal growth conditions for 24 h. Further treatment was as follows:

The dishes were placed on ice and the culture medium was removed. The cells were washed twice in ice-cold PBS. 300 µl of Passive Lysis Buffer (Promega) was added to each dish and left on ice for the cells to lyse for 5 min. Sterile disposable cell scrapers were used to remove the cells from the dish. The cell extract was transferred to a sterile Eppendorf tube and centrifuged for 5 min at 300 rpm in a microcentrifuge. The supernatant was transferred to a fresh sterile Eppendorf tube and stored at -20°C for further analysis (as described in section 2.6.3).

b) Suspension cells

The day before transfection, the cells were seeded so that they were not too dense on the day of transfection. The cells were incubated in their normal culture medium overnight. The day of transfection the cells were centrifuged at 1000 rpm for 5 min. The medium was aspirated and the pellet washed twice in cell culture medium, excluding serum, proteins and antibiotics. While the cells were being washed, 20 µg of DNA (16 µg experimental firefly luciferase DNA and 4 µg *Renilla* luciferase DNA) was pipetted into each of the labeled 4 mm BioRad electroporation cuvettes. The cells were resuspended, in the same culture medium, at a concentration of 2×10^7 cells per ml. 500 µl of suspension cells were added to the DNA in each of the cuvettes and mixed by pipetting up and down three times.

The BioRad electroporation machine was set to 0.975 µF and 250 V. Each cuvette was gently tapped before being placed in the electroporator, to ensure the cells and DNA were mixed. Once electroporated, the cells were incubated on ice for 10 min and then transferred to sterile Eppendorf tubes using a disposable Pasteur pipette, making sure none of the dead cell debris was transferred. The dead cell debris is obvious as it forms a thick layer on the top of the suspension of living cells. The cells were centrifuged for 7 min at 300 rpm in a microcentrifuge. The medium was aspirated, the pellet resuspended in 0.5 ml fresh cell culture medium and transferred to 2.5 ml medium in a 60 mm tissue culture dish. The cells were then returned to their normal growth conditions for 24 h.

The dishes were placed on ice and the culture medium, including the cells, was decanted into 15 ml sterile tissue-culture tubes, on ice. The cells were centrifuged for 5 min at 300 rpm and washed twice in ice-cold PBS. 300 µl of Passive Lysis Buffer (Promega) was added to each dish and left on ice for the cells to lyse for 5 min. The cell extract was transferred to a sterile Eppendorf tube and centrifuged for 5 min at 300 rpm in a microcentrifuge. The supernatant was transferred to a fresh sterile Eppendorf tube and stored at -20°C for further analysis (as described in section 2.6.3).

2.6.2.2 Stable transfection

The cells were transfected in the same manner as for the transient transfections with one exception. The *Renilla* luciferase reporter construct was replaced with a vector encoding a neomycin resistance gene. The cells were allowed to grow until they were 95% confluent and then they were seeded at a number of different dilutions in normal growth medium plus G418 (1 mg/ml). They were incubated for a further 5 days, changing the medium at appropriate intervals. All surviving cell clones were pooled and the number of cells counted. Equal number of cells (to normalize for transfection efficiency) were washed and lysed, as described in section 2.6.2, and then assayed for luciferase activity (as described in section 2.6.3).

2.6.3 Luciferase reporter assay

In the Dual Luciferase Reporter (DLR™) Assay, the activities of firefly (*Photinus pyralis*) and *Renilla* (*Renilla reniformis*) luciferases are measured sequentially from a single sample. The firefly luciferase reporter is measured first by adding Luciferase Assay Reagent II (LAR II) to generate a signal. After quantifying the firefly luminescence, this reaction is quenched, and the *Renilla* luciferase reaction is initiated by simultaneously adding Stop & Glo® Reagent to the same tube. The Stop & Glo® Reagent also produces a luminescent signal from the *Renilla* luciferase.

To prepare the LAR II, 10 ml of luciferase assay buffer was added to the lyophylized luciferase assay substrate, resuspended, dispensed into working aliquots and stored at -80°C. The cells to be assayed were harvested, as above, and 20 µl of each cell extract was added to a well in a 96-well plate. 100 µl of LAR II was added to each well containing cell extract and mixed by pipetting up and down twice.

Firefly luciferase activity was recorded on a MicroBeta workstation (Perkin Elmer). The plate was removed from the machine and 100 µl of Stop & Glo® Reagent added to each sample well. The *Renilla* luciferase activity was recorded using the same apparatus. The transfection efficiency of each sample was calculated from the activity of the *Renilla* luciferase. The experimental luciferase activity was normalized using the promoterless PGL3 reporter vector. The luciferase gene cloned in PGL3 under the control of the mouse p110γ promoter ((174)) or mouse Vav promoter ((124)) were used as positive controls for transient and stable transfections, respectively.

2.7 Rapid Amplification of cDNA ends

Rapid Amplification of cDNA ends (5'-RACE) is a polymerase chain reaction-based technique developed to facilitate the cloning of the 5'-ends of mRNAs. The procedure used specifically selects for capped transcripts. Ambions RLM^{TM} 5'RACE kit was used to identify the 5' transcription start site of the p110δ gene. The manufacturers' protocol was followed.

2.7.1 Materials

First Choice^{TM} RLM-RACE (RNA Ligase Mediated Rapid Amplification of cDNA Ends) kit was purchased from Ambion Ltd, Huntindon, Cambridgeshire.
Phenol:Chloroform: an equal volume of water-saturated phenol and chloroform.
Chloroform: molecular biology grade

2.7.2 5'RACE

10 µg of total RNA, containing the target of interest, was treated with 2 µl Calf Intestinal Phosphatase (CIP), 2 µl 10 x buffer and the sample made up to 20 µl with nuclease-free water. CIP removed the 5'-phosphate from all RNA species except those that had a cap structure (present on all Polymerase II transcripts). Molecules that are dephosphorylated in the CIP reaction include rRNA, tRNA and fragmented mRNA that does not contain the 5'-end (Figure 2-1A).

After 1 h at 37°C the CIP reaction was terminated by phenol:chloroform extraction: 15 µl ammonium acetate, 115 µl dH_2O and 150 µl acid phenol:chloroform were vortexed and centrifuged for 5 min at maximum speed in a microcentrifuge. The aqueous layer was transferred to a sterile tube and 150 µl of isopropanol was mixed with the sample by vortexing. The tube was stored on ice for 10 min and then centrifuged at maximum speed for 20 min. The supernatant was discarded and the pellet washed in 0.5 ml ice-cold 70% ethanol. The tube was re-centrifuged at maximum speed for a further 5 min, the ethanol removed and the pellet allowed to air dry. The RNA pellet was resuspended in 11 µl nuclease-free water.

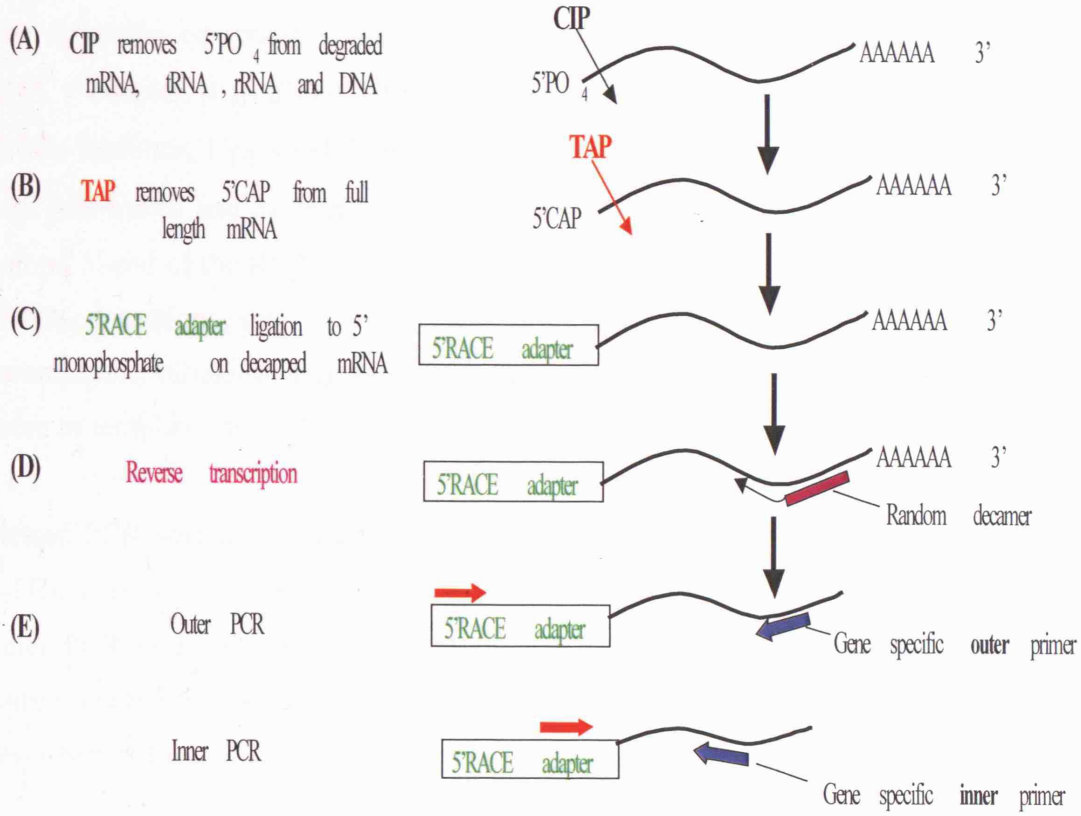

Figure 2-1 FirstChoice™ RLM-RACE Protocol

Tobacco Acid Pyrophosphatase (TAP) was used to remove the cap structure from mRNA, leaving a 5'-phosphate (Figure 2-1B). 5 µl of the CIP-treated RNA was mixed with 1 µl 10 x TAP buffer, 2 µl TAP and 2 µl nuclease-free water. The sample was incubated at 37°C for 1 h.

Next a synthetic RNA oligonucleotide, of known sequence, was added and ligated to the CIP/TAP-treated RNA (Figure 2-1C). 2 µl of CIP/TAP-treated RNA was mixed with 1 µl 5'RACE adapter, 1 µl 10 x ligase buffer, 2 µl T4 RNA ligase (2.5 U/µl) and 4 µl nuclease-free water. The mixture was incubated for 1 h at 37°C. This procedure is aimed at adding an adapter to the decapped mRNA only, and not to the dephosphorylated molecules.

Sequence of the RACE adapter:
5'GCUGAUGGCGAUGAAUGAACACUGCGUUUGCUGGCUUUGAUGAAA**3'**

The RNA was reverse transcribed using random decamers as primers (Figure 2-1D). The following components were assembled in an RNase-free microfuge tube: 2 µl RNA + adapter, 4 µl dNTP MIX, 2 µl random decamers, 2 µl 10 x RT buffer, 1 µl RNase Inhibitor, 1 µl M-MLV reverse transcriptase and 6 µl nuclease-free water. The tube was mixed and incubated at 42°C for 1 h. When the RT reaction extends to the natural 5'-end of the RNA, it will incorporate the adapter sequence into the first strand cDNA. Reactions that did not incorporate the adapter sequence (e.g. due to incomplete synthesis or lack of an adapter on the template RNA) are not expected to serve as template for PCR amplification.

Nested PCR was performed using gene-specific primers and adapter primers (Figure 2-1E). 1 µl of the RT reaction was used for the outer RLM-RACE PCR. 1 µl of the outer PCR reaction was used for the inner 5'RLM-RACE PCR. The PCR reactions were prepared following the protocol described in section 2.3.2. The primers used are described in Table 2-2. The PCR cycling profile was as follows:

94°C for 3 min		x 1 cycle
94°C for 30 sec	}	x 35 cycles
60°C for 30 sec		
72°C for 30 sec		
72°C for 7 min		x 1 cycle

The PCR products were run on a 2% agarose gel as described in section 2.3.3.1 and then eluted using the QIAquick method (as described in section 2.3.4). The Taq DNA polymerase used in the PCR reaction ensured easy cloning of the PCR product into the pGEM®-T Easy vector as described in section 2.4.3a. The samples were prepared for DNA sequencing, as described in section 2.5 and the sequence analysed using The Blast Like Alignment Tool (BLAT) (186) (section 2.11).

2.8 Real-time reverse transcriptase (RT) PCR

2.8.1 Materials

TaqMan® Universal PCR mastermix: (Catalogue number: 4304437) Applied Biosystems.

The primers and probes used were all obtained from Applied Biosystems as part of the "TaqMan Gene Expression Assays":

Gene Symbol	ABI assay ID	Gene name
β-actin	Mm00607939_s1	β-actin mouse
β-ACTIN	Hs99999903_m1	β-actin human
cd45	Mm00448463_m1	cd45 mouse
pik3ca	Mm00435669_m1	PI3K p110α
pik3cd	Mm00435674_m1	PI3K p110δ mouse
PIK3CD	Hs00192399_m1	PI3K p110δ human
pik3cd -2a	Assay-on-demand	PI3K p110δ detects -2a transcripts mouse
pik3cd -2b	Assay-on-demand	PI3K p110δ detects -2b transcripts mouse
PIK3CD -2a	Assay-on-demand	PI3K p110δ detects -2a transcripts human
PIK3CD -2b	Assay-on-demand	PI3K p110δ detects -2b transcripts human

2.8.2 Method

Real-Time chemistries allow for the detection of PCR amplification during the early phases of the reaction. Measuring the kinetics of the reaction in the early phases of PCR provides a distinct advantage over traditional PCR detection. Traditional methods use agarose gels for detection of PCR amplification at the final phase of the PCR reaction.

A basic PCR run can be broken up into three phases (Figure 2-2):

Exponential: This is when there is an exact doubling of the product and the reaction is very specific and precise.

Linear: The reaction components are being consumed, the reaction is slowing and the products are starting to degrade.

Plateau (end-point: gel detection for traditional methods): the reaction has stopped, no more PCR products are being made and if left long enough the PCR products will begin to degrade.

Measurements at the plateau phase, as is done in traditional PCR techniques would not truly represent the initial amount of starting material.

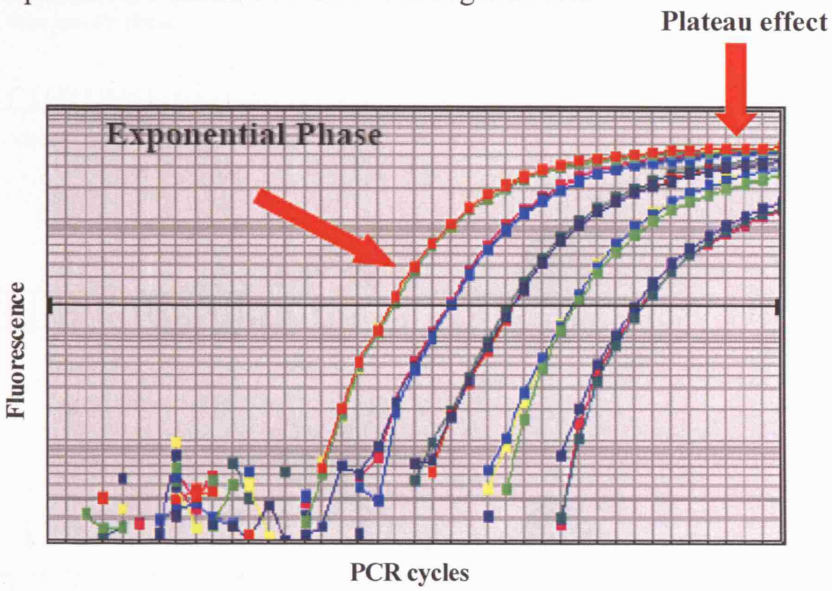

Figure 2-2 Log view of a 5-fold dilution series. All of the samples seem to plateau at the same place even though the exponential phase shows a clear difference between the dilution series.

Applied Biosystems has designed and manufactured nearly 40,000 real-time PCR assays for measuring the expression of genes in humans, mice and rats. TaqMan Assay-on-Demand Gene Expression Products each consist of amplification primers (forward and reverse) and a fluorescent labeled TaqMan® probe all in a single tube.

The TaqMan® probe is designed to anneal to a complementary sequence of template between the forward and reverse primers. The probe sits in the path of the of the polymerase enzyme as it starts to copy the cDNA. When the enzyme reaches the annealed probe the 5' exo-nuclease activity of the enzyme cleaves the probe (Figure 2-3).

The TaqMan® probe has a high-energy dye, a reporter, at the 5' end, and a low energy molecule, the quencher, at the 3'-end. When the probe is intact and excited by a light source, the emission of the Reporter dye is suppressed by the Quencher dye as a result of the close proximity of the dyes. When the probe is cleaved, the Reporter and the

Quencher separate causing the transfer of energy to stop, leading to an increase in the fluorescent emissions of the reporter.

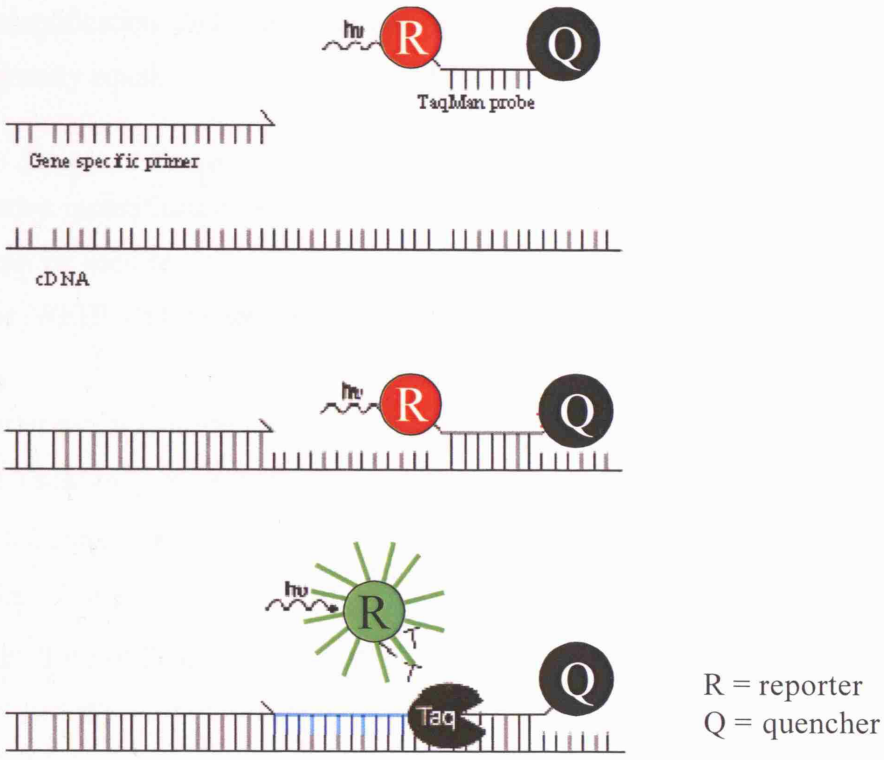

Figure 2-3 A diagrammatic representation of the chemistry behind the ABI Real-Time PCR protocol (reverse primer not shown).

The ABI PRISM 7700 Sequence Detection System detects the reporter signal. The amount of reporter signal is proportional to the amount of product being produced for a given sample.

Reactions are characterised by the point in time (or PCR cycle) where the target amplification is first detected. This value is referred to as the cycle threshold (C_T), the cycle at which fluorescence intensity is greater than background fluorescence. Consequently the greater the quantity of target mRNA in the starting material, the faster a significant increase in fluorescent signal will appear, yielding a lower C_T.

The method of relative quantitation of mRNA was used to determine the mRNA expression levels of the p110δ gene in different cell lines and tissue samples. Using this method the quantity of mRNA in the experimental sample is expressed relative to the quantity of mRNA in a reference sample, the calibrator.

There are two methods of quantification, namely the standard curve method and the comparative C_T method ($\Delta\Delta CT$). For this method to be valid, the efficiency of the target amplification and the efficiency of amplification of the reference must be approximately equal.

2.8.2.1 Standard Curve

For relative quantification, stock of any mRNA or cDNA containing the appropriate target can be used to prepare the standards. In this thesis, we have been using cDNA from the WEHI 231 mouse B cell line and the Jurkat T cell lymphoma human cell line.

One master mix was made for each primer set used:

10 µl of TaqMan® Universal PCR mix (polymerase and buffer),

1 µl 20 x Gene expression assay mix (primers and probe)

8 µl RNase free-water

Example of a relative standard curve:

19 µl of the appropriate master mix, prepared as described above, was added to the wells containing 1 µl of sample cDNA. The concentration of the mRNA samples is calculated so that an equal concentration is added to all the wells.

Table 2-3 shows the plate set-up for the relative quantitation of the p110δ mRNA where the target and endogenous reference are amplified in separate wells. Rows A-D contains p110δ-specific primers and a p110δ labelled probe. Table 2-4 shows the plate set-up for GAPDH mRNA. Rows E-H contains GAPDH-specific primers and a GAPDH labelled probe.

	1	2	3	4	5	6	7	8	9	10	11	12
A	NTC	NTC	NTC	STND 1000 pg	STND 1000 pg	STND 1000 pg	STND 500 pg	STND 500 pg	STND 500 pg	STND 500 pg	STND 500 pg	STND 500 pg
B	STND 100 pg	STND 100 pg	STND 100 pg	STND 50 pg	STND 50 pg	STND 50 pg	STND 20 pg	STND 20 pg	STND 20 pg	STND 10 pg	STND 10 pg	STND 10 pg
C	p110δ NIH3T3	p110δ NIH3T3	p110δ NIH3T3	p110δ NIH3T3	p110δ NIH3T3	p110δ NIH3T3	p110δ LLC	p110δ LLC	p110δ LLC	p110δ LLC	p110δ LLC	p110δ LLC
D	p110δ A20	p110δ A20	p110δ A20	p110δ A20	p110δ A20	p110δ A20	p110δ EL4	p110δ EL4	p110δ EL4	p110δ EL4	p110δ EL4	p110δ EL4
E												
F												
G												
H												

Table 2-3 A table showing the 96-well plate set up for standard curve method for relative quantitation of the p110δ mRNA. WEHI 231 cDNA was used for the standard (**STND**). Wells A1 – A3 were used as no template controls (**NTC**).

	1	2	3	4	5	6	7	8	9	10	11	12
A												
B												
C												
D												
E	NTC	NTC	NTC	STND 1000 pg	STND 1000 pg	STND 1000 pg	STND 500 pg	STND 500 pg	STND 500 pg	STND 500 pg	STND 500 pg	STND 500 pg
F	STND 100 pg	STND 100 pg	STND 100 pg	STND 50 pg	STND 50 pg	STND 50 pg	STND 20 pg	STND 20 pg	STND 20 pg	STND 10 pg	STND 10 pg	STND 10 pg
G	GAPDH NIH3T3	GAPDH NIH3T3	GAPDH NIH3T3	GAPDH NIH3T3	GAPDH NIH3T3	GAPDH NIH3T3	GAPDH LLC	GAPDH LLC	GAPDH LLC	GAPDH LLC	GAPDH LLC	GAPDH LLC
H	GAPDH A20	GAPDH A20	GAPDH A20	GAPDH A20	GAPDH A20	GAPDH A20	GAPDH EL4	GAPDH EL4	GAPDH EL4	GAPDH EL4	GAPDH EL4	GAPDH EL4

Table 2-4 A table showing the 96-well plate set up for the endogenous GAPDH mRNA control. WEHI 231 cDNA was used for the standard (**STND**). Wells E1 – E3 were used as no template controls (**NTC**).

Constructing a relative standard curve:

The C_T values can be exported from the ABI prism 7700 Sequence detection software to an Excel spreadsheet. From here a standard curve can be produced:

1) Calculate the log of the ng of each standard sample and add it to the spreadsheet as shown below in Table 2-5:

Sample	ng	log ng	Ct
NTC			40
NTC			40
NTC			40
WEHI 231-1	1	0	25.63
WEHI 231-1	1	0	25.56
WEHI 231-1	1	0	25.59
WEHI 231-0.5	0.5	-0.30103	26.83
WEHI 231-0.5	0.5	-0.30103	26.8
WEHI 231-0.5	0.5	-0.30103	26.67
WEHI 231-0.2	0.2	-0.69897	28.2
WEHI 231-0.2	0.2	-0.69897	28.13
WEHI 231-0.2	0.2	-0.69897	28.1
WEHI 231-0.1	0.1	-1	29.29
WEHI 231-0.1	0.1	-1	29.2
WEHI 231-0.1	0.1	-1	29.04
WEHI 231-0.05	0.05	-1.30103	30.14
WEHI 231-0.05	0.05	-1.30103	30.11
WEHI 231-0.05	0.05	-1.30103	30.17
WEHI 231-0.02	0.02	-1.69897	31.35
WEHI 231-0.02	0.02	-1.69897	31.34
WEHI 231-0.02	0.02	-1.69897	31.63
WEHI 231-0.01	0.01	-2	32.55
WEHI 231-0.01	0.01	-2	32.33
WEHI 231-0.01	0.01	-2	32.37

Table 2-5 An Excel spreadsheet showing the CT values obtained for the different concentrations of WEH 231 mRNA

2) Using Excel chart wizard, draw an XY scatter plot with the log input as the X values and the CT as the Y values

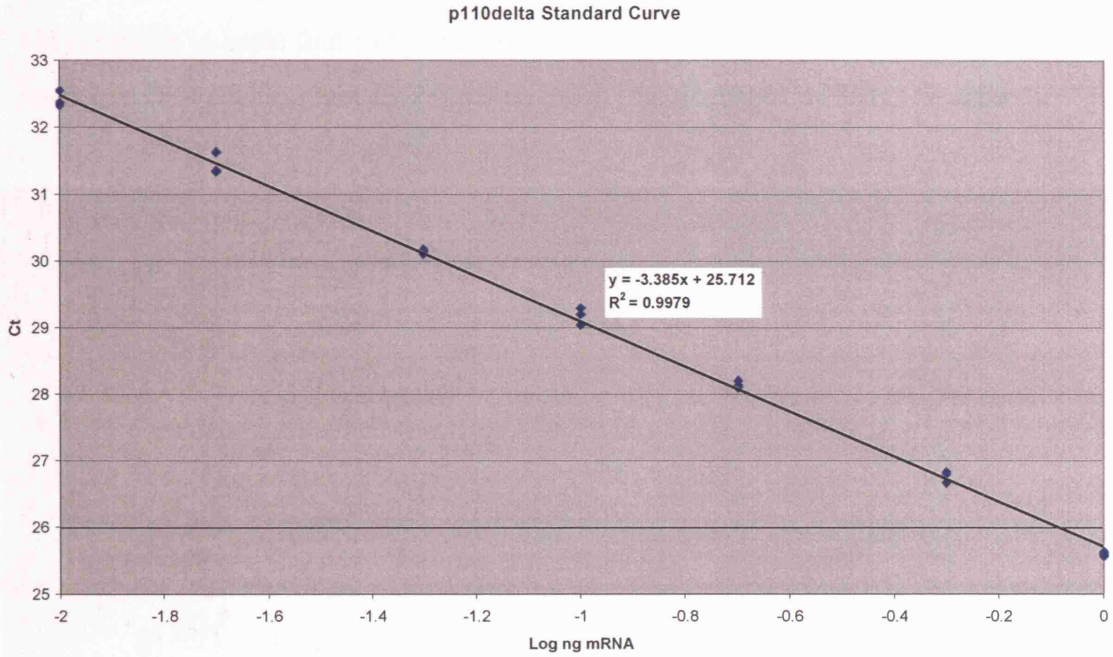

Figure 2-4 The standard curve for the amplification of p110δ target in WEHI 231 mRNA

The next step is to calculate the input amount of the unknown samples.

Calculating the input amount:

1) Calculate the log input amount for each sample by dividing (C_T-b) by m

Where b is the y intercept of the standard curve line, and m = the slope of the standard curve. In the graph above y = 25.712 and m = -3.385

2) To calculate the input amount the log input must be inversed: $10^{(\log \text{input})}$.

This calculation must be repeated for all of the samples from the p110δ-primed wells and the GAPDH-primed wells (data not shown). The average amount of p110δ mRNA must be divided by the average amount of GAPDH mRNA to determine the normalised amount of p110δ ($p110\delta_N$). An example is shown in Table 2-6.

Comparing samples with a calibrator:

The normalised amount of target ($p110\delta_N$) is a unitless number that can be used to compare the relative amount of target in different samples. One way to do this is to designate one of the samples as a calibrator. Usually the sample with the lowest expression is used as the calibrator, so in our case we chose NIH3T3.

Each p110δ_N value in Table 2-6 below is divided by the NIH3T3 p110δ_N value to give the values of p110δ in each cell line relative to p110δ in the NIH3T3 cell line. These results indicate that LLC cells contain 5.5x as much, A20 cells contain 34.2x as much and EL4 cells contain 15.7 times as much p110δ mRNA as NIH3T3 cells.

Cell line	p110δ ng WEHI 231 RNA	GAPDH ng WEHI 231 RNA	p110δ_N Norm. to GAPDH	p110δ_N relative to NIH3T3
NIH 3T3	0.033	0.51		
	0.043	0.56		
	0.036	0.59		
	0.043	0.53		
	0.039	0.51		
	0.040	0.52		
Average	0.039	0.54	0.07	1.0
LLC	0.40	0.96		
	0.41	1.06		
	0.41	1.05		
	0.39	1.07		
	0.42	1.06		
	0.43	0.96		
Average	0.41	1.02	0.40	5.5
A20	0.67	0.29		
	0.66	0.28		
	0.70	0.28		
	0.76	0.29		
	0.70	0.26		
	0.68	0.27		
Average	0.70	0.28	2.49	34.2
EL4	0.97	0.82		
	0.92	0.88		
	0.86	0.78		
	0.89	0.77		
	0.94	0.79		
	0.97	0.80		
Average	0.93	0.81	1.15	15.7

Table 2-6 Amounts of p110δ and GAPDH in mouse NIH 3T3, LLC, A20 and EL4 cell lines

2.8.2.2 $\Delta\Delta C_T$ method

The comparative C_T ($\Delta\Delta C_T$) method is similar to the standard curve method, except it uses arithmetic formulas to achieve the same result for relative quantitation. It is possible to eliminate the use of standard curves for relative quantitation as long as a validation experiment is performed.

Validation experiment:

For the $\Delta\Delta C_T$ calculation to be valid, the efficiency of the target amplification and the efficiency of the reference amplification must be approximately equal. For this to be assessed one must see how ΔC_T varies with dilution. The data from the standard curve used in the previous example will be used to demonstrate the calculations necessary to prepare a relative efficiency plot.

Input amount ng mRNA	p110δ Average C_T	GAPDH Average C_T	ΔC_T p110δ - GAPDH
1	25.59	22.64	2.95
0.5	26.77	23.73	3.04
0.2	28.14	25.12	3.02
0.1	29.18	26.16	3.01
0.05	30.14	27.17	2.97
0.02	31.44	28.62	2.82
0.01	32.42	29.45	2.97

Table 2-7 Average CT value for p110δ and GAPDH at different input amounts

If the efficiencies of the two amplicons are approximately equal, the plot of log input versus ΔC_T has a slope of approximately zero, as shown in Figure 2-5.

The absolute value of the slope of log input amount versus ΔCT should be less than 0.1. The slope in Figure 2-5 is 0.0492, which passes this test. Once this is proven the ΔΔCT calculation can be used for the relative quantitation of target without running standard curves on the same plate.

If the efficiency of the two primers is not equal quantification must be performed using the standard curve method.

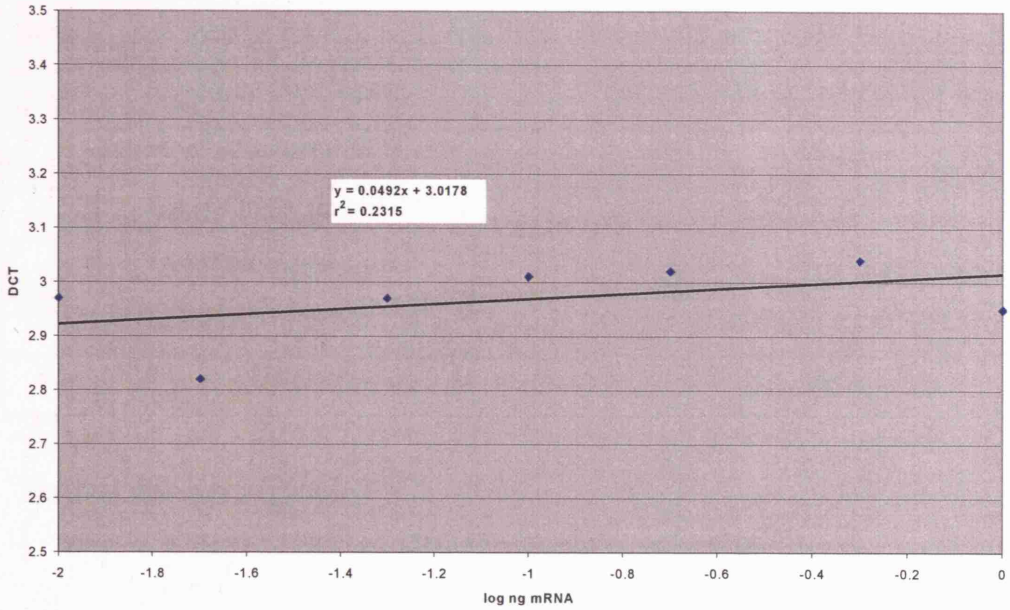

Figure 2-5 Plot of log input amount versus ΔC_T

Comparative C_T results

The C_T data used to determine the amounts of p110δ (Table 2-5) and GAPDH mRNA (data not shown) will be used to illustrate the $\Delta\Delta C_T$ calculation. Table 2-8 shows the average C_T results for the mouse NIH3T3, LLC, A20 and EL4 cell line samples and how these C_T results are used to determine ΔC_T, $\Delta\Delta C_T$ and the relative amount of p110δ mRNA.

Cell line	p110δ Average C_T	GAPDH Average C_T	ΔC_T p110δ-GAPDH[a]	$\Delta\Delta C_T$ ΔC_T-$\Delta C_{T\ NIH3T3}$[b]	$p110\delta_N$ Relative to NIH3T3[c]
NIH3T3	30.49	23.63	6.86	0.00	1.0
LLC	27.03	22.66	4.37	-2.50	5.6
A20	26.25	24.60	1.65	-5.21	37.0
EL4	25.83	23.01	2.81	-4.05	16.5

Table 2-8 Relative quantitation using the comparative C_T method

a. The ΔC_T value is determined by subtracting the average GAPDH C_T value from the average p110δ C_T value.
b. The calculation of $\Delta\Delta C_T$ involves subtraction by the ΔC_T calibrator value.
c. $p110\delta_N$ is calculated using the equation $2^{-\Delta\Delta CT}$.

2.8.2.3 Basic protocol

The total RNA was isolated as described in section 2.4.4.3. It was reverse transcribed using the method described in section 2.3.2b. Concentration was determined using the method described in 2.4.4.5. PCR reactions were set up as shown in the table below:

Reaction Component	Volume/Well (20 µl volume reaction)
TaqMan® Universal PCR mix	10 µl
20 x Assay-On-Demand™ Gene expression assay mix	1 µl
cDNA in RNase-free water	9 µl (0.25 µg cDNA)
Total	20 µl

Thermal cycler conditions

95°C for 10 min x 1 cycle

95°C for 15 sec
60°C for 1 min } x 40 cycles

The CT values were exported into an Excel file and the relative values calculated as shown above.

2.9 RNase protection

2.9.1 Materials

MAXIscript™: *In vitro* transcription kit was purchased from Ambion Ltd, Huntindon, Cambridgeshire.

RPA III™ (Ribonuclease Protection Assay) was purchased from Ambion Ltd, Huntindon, Cambridgeshire.

Denaturing gel (22.5 ml): 10.8 g of urea, 2.25 ml 10 x TBE, 2.85 ml 40% Acrylamide : bis-acrylamide made up to 22.5 ml with dH_2O

10x TBE: 0.9 M Tris base, 0.9 M boric acid, 20 mM 0.5 M EDTA solution

SequaGel6: ready-to-use 19:1 denaturing DNA gel solution (National Diagnostics) contains 6% acrylamide, 1X TBE (89 mM Tris Base, 89 mM Boric Acid, 2 mM EDTA, pH 8.3) and 6 M Urea

[α^{32}P]UTP ~ 800 Ci/mmol

2.9.2 Method

The Ribonuclease Protection assay (RPA) is an sensitive procedure for the detection and quantitation of individual RNA species in a sample of total RNA or poly(A) RNA. A radioactively labeled RNA probe is synthesized that is complementary to part of the target RNA to be analysed. The labeled probe and sample RNA are hybridized and the mixture treated with RNase A to degrade the unhybridized probe. The labeled probe hybridizes to complementary RNA from the sample is protected from RNase digestion and is separated on a polyacrylamide gel and visualized by autoradiography. If the probe is present in excess the intensity of the protected fragment is directly proportional to the amount of target RNA in the sample.

2.9.2.1 Probe preparation

RNA transcripts are useful as probes in hybridisation reactions because RNA-RNA and RNA-DNA hybrids are more stable than DNA-DNA duplexes. For the RNase protection assay ^{32}P-labelled UTP was incorporated into the RNA transcript.

The 5'RACE p110δ PCR products, cloned into the pGEM®-T Easy Vector were used as templates to create antisense probes to p110δ mRNA. The vector was linearised,

using the *Sph I* restriction enzyme site, at the 5' end of the insert to prevent the transcription of the vector (method described in section 2.3.3). The pGEM®-T Easy Vector has a T7 and SP6 RNA polymerase promoter at either end of the cloning site (Figure 2-6).

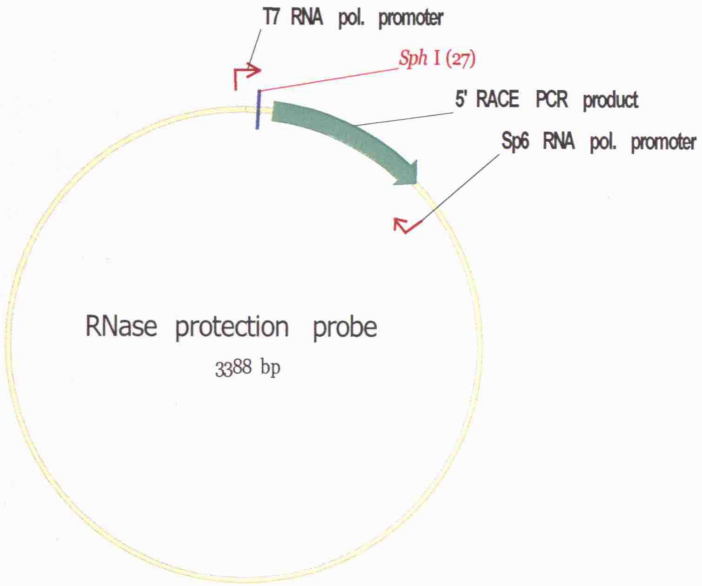

Figure 2-6 The pGEM®-T Easy vector. The orientation of the PCR insert is shown by the direction of the arrow. The polymerase promoter sites and the direction in which they transcribe, are identified by the smaller arrows. The *Sph I* restriction enzyme site, used to linearise the vector is shown close to the T7 polymerase promoter site at the 5' end of the insert.

After linearization the DNA was precipitated: 2 µl 3 M sodium acetate and 40 µl 100% EtOH were added to the sample, mixed and chilled at -20°C for 15 min. The sample was pelleted by centrifugation at full speed for 15 min. The supernatant was removed and the sample respun to remove all of the liquid. The pellet was resuspended to a final volume of 0.5 µg/ µl.

The following components were assembled at room temperature for the RNA transcription reaction: 1 µg vector DNA, 2 µl 10 x transcription buffer, 1 µl 10 mM ATP, 1 µl 10 mM CTP, 1 µl 10 mM GTP, 5 µl ^{32}P-UTP and 2 µl SP6 RNA polymerase. The volume was made up to 20 µl with nuclease-free water. The sample was mixed thoroughly and incubated at 37°C for 1 h. 1 µl DNase I was incubated at 37°C for 15 min with sample, to remove the DNA template.

After the transcription reaction the probe was gel-purified on a 5% denaturing acrylamide gel: 10.8 g of urea, 2.25 ml 10 x TBE, 2.85 ml 40% Acrylamide:bis-

acrylamide made up to 22.5 ml with dH_2O. These components were mixed on a magnetic stirrer until the urea had dissolved completely. Immediately before the gel was poured 120 µl of 10% APS and 16 µl of TEMED were added to the gel solution. The well-forming comb was put in place and the gel allowed to set.

The wells of the gel were washed thoroughly and the sample loaded. Following separation by gel electrophoresis at 250 V for 3 h in 1 x TBE the gel was exposed to X-Ray film. The film and the gel were aligned and the position of the probe was marked on the gel. Using a sterile scalpel and forceps the radioactive probe was cut out of the gel and transferred to a RNase-free microcentrifuge tube. 350 µl of probe elution buffer was added to the tube and incubated at 37°C overnight. The activity, in counts per minute (cpm), of the labelled probes was calculated using a scintillation counter.

2.9.2.2 Standard hybridization procedure

50,000 cpm of each probe was mixed with 10 µg total RNA. Probe and RNA were concentrated using at speedvac vacuum centrifuge until there was less than 4 µl in each sample. 10 µl of hybridization buffer was used to resuspend the probe/RNA samples and the combination was heated to 95°C for 5 min. The samples were vortexed before being incubated at 45°C in a waterbath overnight to allow hybridization.

2.9.2.3 RNase digestion of hybridized probe and sample RNA

Two large DNA sequencing gel glass plates were cleaned thoroughly and the smaller of the two was siliconised to easy the separation of the plates after electrophoresis. 100 µl of APS and 10 µl of TEMED was added to 70 ml of prepared SequaGel 6. The gel was poured and allowed to set.

150 µl of RNase/digestion buffer mix was added to each of the hybridized samples. Buffer alone was added to the controls which lacked RNase. The samples were vortexed and centrifuged before being incubated at 37°C for 30 min. 225 µl of RNase Inactivation/precipitation III solution was then added followed by vortexing and

centrifugation of the samples. A second incubation step at –20 °C for at least 15 min preceded a centrifugation at maximum speed for 15 min.

The supernatant was carefully discarded and the pellet was resuspended in 10 µl Gel loading buffer II. The samples were incubated for 3 min at 95 °C before being stored on ice until ready to load on a gel.

Following separation by gel electrophoresis at 250 V for 3 h in 1 x TBE the gel was exposed to a BioRad phosphoimaging screen overnight.

2.10 Protein Analysis

2.10.1 Materials

SDS-PAGE MW (molecular weight) marker was obtained from Sigma Aldrich and contained proteins of the following molecular weight: 180,000, 116,000, 84,000, 58,000, 48,500, 36,500, 26,600 kDa

ECL (enhanced chemiluminescence) kit was purchased from Amersham Biosciences UK Ltd, Buckinghamshire

Bio-Rad Protein Assay was purchased from Bio-Rad Laboratories Ltd., Hemel Hempstead, Hertfordshire.

Immobilon P: Polyvinylidene fluoride (PVDF) membrane was purchased from Millipore.

Cell Lysis Buffer Stock: 50 mM Tris.HCl pH 7.4, 1 mM EDTA, 150 mM NaCl, 1% w/v Triton X-100.

Inhibitors: 1000x sodium fluoride from 1 M NaF stock (stored at -20°C); 100x sodium orthovanadate from 100 mM stock (stored at -20°C); 100x aprotinin from 5 TIU/mg (trypsin inhibitor units/mg) (stored at 4°C); 1000x leupeptin from 10 mM stock (stored at -20°C); 1000x pepstatin A, dissolved in DMSO to 0.7 mg/ml (stored at -20°C); 1000x TLCK (Na-p-tosyl-L-lysine chloromethyl ketone hydrochloride) dissolve to 10 mg/ml in 1 mM HCl (stored at -20°C); 1000x DTT (dithiothreitol) from 1 M stock (stored at -20°C); 100x PMSF (phenylmethylsulphonyl fluoride) from 100 mM stock in isopropanol (stored at -20°C).

Blocking Buffer: 5% (w/v) skimmed milk powder in TBS-T

Resolving gel:

Percentage gel	8%	10%
30% acrylamide stock (bis-acrylamide ratio - 37.5:1)	10 ml	13.7 ml
4x SEP buffer	10 ml	10 ml
dH_2O	18.8 ml	16 ml
10% APS	0.25 ml	0.25 ml
TEMED	0.025 ml.	0.025 ml.

5x Running buffer: 0.96 M glycine, 0.125 M Tris, 0.5% (w/v)

5x Sample buffer: 156.25 mM Tris.HCl pH6.8, 5% (w/v) SDS, 25% (v/v) glycerol, 0.0025% (w/v) bromophenol blue

4x SEP: 1.5 M Tris.HCl pH 8.8, 0.4% (w/v) SDS

4x SPAC: 0.5 M Tris.HCl pH 6.8, 0.4% (w/v) SDS

Stacking gel (4%): 30% acrylamide stock (bis-acrylamide ratio - 37.5:1) - 2.7 ml, 4x SPAC buffer-5 ml, dH_2O-12.2 ml, 10%APS-0.9 ml, TEMED-0.025 ml.

TBS (Tris-Buffered Saline): 20 mM Tris.HCl pH 7.5, 137 mM NaCl in ddH_2O

TBS-T (TBS with tween): 0.1% (v/v) Tween 20 in TBS

Wet Transfer buffer: 58 g glycine, 11.6 g Tris, 2 g SDS, 400 ml methanol made up to 2 L with ddH_2O

PDGFR peptide: pYVPMLG (Y = phosphotyrosine) purchased from Alta Bioscience, Edgbaston, Birmingham.

Actigel kit: purchased from Sterogene Bioseparations,. Carlsbad, CA

0.1 M Phosphate buffer: 8.5 ml 0.2 M NaH_2PO_4, 91.5 ml 0.2 M $NaHPO_4$, 100 ml dH_2O.

Coupled peptide storage solution: 8 ml 50 mM Tris.HCl pH 8.0, 80 µl $NaVO_3$, 20 µl NaN_3 per 10 ml.

Primary antibodies and concentrations used for western blot

p110δ: Rabbit polyclonal. Made in-house and used at 1 in 5000 dilution. Immunogen: the last 17aa of the human p110δ protein.

p110δ: Rat monoclonal (Transduction Laboratories; P75220) – used at 1:500 dilution. Immunogen: amino acids 73-90 in the N terminal region of the mouse p110δ protein.

pan p85: Rabbit polyclonal (Upstate Group LLC; 06-195) - used at 1:5000 dilution. Immunogen: the N-terminal SH2 domain of rat p85 protein.

CD45: Rat monoclonal, (BD Pharmingen; clone 30-F11) - used at 1 in 1000 dilution

β-actin: Mouse monoclonal (Sigma; clone AC-15) – used at 1 in 5000 dilution

GAPDH: Mouse monoclonal (Abcam, Cambridge; 6CS) – used at 1 in 5000 dilution

Secondary antibodies

Anti-mouse Ig: HRP-conjugated sheep (Amersham Pharmacia) – used at 1 in 5000

Anti-rabbit Ig: HRP-conjugated donkey (Amersham Pharmacia) – used at 1 in 5000

Anti-rat Ig: HRP-conjugated goat (Amersham Pharmacia) – used at 1 in 5000

2.10.2 Western Blots

Cell lysis was always carried out on sub-confluent cells (70-80% confluency). Cells were washed twice in ice-cold PBS and lysed in lysis buffer supplemented with protease and phosphatase inhibitors. Cells from a 10 cm plate were normally lysed in 500 µl of lysis buffer and scraped on ice. Lysates were left 20 min on ice and cell debris was cleared by centrifugation at 14,000 rpm for 10 min at 4°C.

Protein concentrations were determined using the Bio-Rad Protein Assay. 5 µl of total cell lysate was mixed with 1 ml of Coomassie protein assay reagent (BioRad). Absorbance readings were taken at 595 nm in a spectrophotometer. Cell lystates were then reduced and denatured in sample buffer and boiled for 5 min at 100°C.

For western blotting, 50 -100 µg of protein lysate was separated by SDS-PAGE using standard procedures. The percentage of the resolving gel was adjusted according to the molecular weight of the protein of interest (see materials for components and their volumes of 8% and 10% resolving gels). All stacking gels were 4% acrylamide.

Following separation by gel electrophoresis, lysates were electroblotted onto polyvinylidene fluoride (PVDF) membranes using a wet transfer tank in transfer buffer. Membranes were then blocked for at least 1 h with 5% (w/v) low fat milk in Tris-buffered saline with 0.05% Tween-20 (TBS-T). Membranes were incubated for 4 h in a primary antibody solution diluted in TBS-T. Membranes were washed in TBS-T (three times 10 min) and then probed with the appropriate horseradish peroxidase (HRP)-coupled secondary antibody. After further washes in TBS-T, protein bands were visualised using the enhanced chemiluminescence (ECL) method.

2.10.3 Immunoprecipitation

Cells were lysed in lysis buffer as described in section 2.10.2. 500 µg of total cellular lysate was used for immunoprecipitations (IP). After spinning the cell lysates for 10 min, the cleared supernatant was collected and incubated with 2 µg of primary antibody and protein-A coupled to sepharose beads for 2 h at 4°C on a rotating wheel. A small amount of supernatant was collected prior to incubation with beads and

antibody to allow the measurement of protein expression in total cell lysate in the same samples used for IP. Immunocomplexes were washed 5 times in lysis buffer, re-suspended in 60 µl sample buffer and boiled for 5 min at 100°C prior to separation by SDS-PAGE. Immunoprecipitates were then analysed by western blotting with the appropriate antibody.

2.10.4 'PDGF receptor peptide bead' pull down assay

For PDGF receptor pull-down experiments, cells were processed similarly to the immunoprecipitation protocol; except that cells were incubated for 1 hr with 2 µg of phosphotyrosine peptide (mimicking the tyrosine phosphorylated part of the PDGF receptor) immobilized on Actigel ALD sepharose beads (Sterogene). Proteins pulled out using this method were then analysed by western blotting with the appropriate antibody.

2.10.4.1 Coupling PDGF receptor peptides

8 mg of peptide were dissolved in 5 ml 0.1 M phosphate buffer in a 15 ml tissue-culture tube. 4 ml of packed actigel beads and 1 ml ALD coupling solution were added to the dissolved peptide. The tube was placed on a wheel at 4°C to mix overnight. The following morning the beads were transferred to a 50 ml tissue-culture tube and washed twice in 40 ml 100 mM Tris.HCl pH 8 + 500 mM NaCl and once in 40 ml 100 mM Tris.HCl pH 8 alone. The beads were centrifuged at 2,500 rpm for 10 min and the supernatant discarded. They were re-suspended in 100 mM Tris.HCl pH 8 on incubated on a wheel for 4 h. The amines present in the Tris interacted with and blocked any uncoupled sites. The beads were centrifuged at 2,500 rpm for 10 min, the supernatant discarded and they were then resuspended in peptide storage solution and stored at 4°C.

2.11 Bioinformatics

2.11.1 NCBI

The human and mouse p110δ genomic sequence was extracted from the NCBI website: **http://www.ncbi.nlm.nih.gov/**. The Human Genome Resources and the Mouse Genome Resources were selected to find the human and mouse p110δ genomic sequences, respectively. PIC3CD was entered as the search parameter. Selecting the 'dl' option, highlighted after the gene name on the chromosome, displayed the sequence. The area displayed was adjusted to include 8 kb upstream of exon -2b in both the mouse and human gene. Altering the numeric information in the 'FROM and TO' box and then selecting display achieved this. The sequence was copied and pasted into Vector NTI and annotated with the recently identified and previously published exon and intron information.

2.11.2 BLAT

Identification and analysis of published ESTs was achieved using the Blast Like Alignment Tool (BLAT) – **http://www.genome.cse.ucsc.edu**. Full-length sequences of the mouse and human p110δ gene were pasted into the search engine on this site.
The software shows alignments between the submitted sequence, human/mouse mRNAs/ ESTs in GenBank® and the genome. GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (187). There are approximately 41,893,844,733 bases in 37,549,400 sequence records as of August 2004. Just over the past year the number of ESTs in the database has increased by 29% to a total of 23.4 million, 4.2 million of those belong to *M. musculus* and 5.7 million to *H. sapiens*.

2.11.3 LAGAN

LAGAN pairwise alignment: **http://lagan.stanford.edu/lagan_web/index.shtml** was used to identify regions of homology between the mouse and human p110δ gene. To submit a sequence to LAGAN it must be in FASTA format. To convert a sequence to

FASTA format it must be pasted into NOTEPAD. The first line of the sequence must begin with >. This can be followed by a definition of the sequence. For example:

>full length human p110d

ATGGGCCAAGCTAATTTTGGGAGGAATTTAGTTTATGGTTTAACCTTAAA

GCGAGGATGAAAATAACCCTTCCCAAAACTAAACCGGATTGGTAAGACTA

.....

LAGAN is a pairwise global alignment program that allows the user to submit their own parameters. It takes local alignments generated by CHAOS as anchors. CHAOS is a pairwise local aligner optimised for non-coding, and other poorly conserved regions of the genome. It uses both exact matching and degenerate bases, and is able to find homology in the presence of gaps.

The LAGAN server has been combined with the VISTA server to produce alignments and visualizations for all requests. LAGAN/VISTA submission forms are available at **http://genome.lbl.gov/cgi-bin/VistaInput?align_pgm=lagan&num_seqs=2** where sequences are submitted.

Sequences can be annotated. The user must submit an annotation file in FASTA format. This consists of a list of the position of each 5'UTR or exon on the sequence submitted. For example:

```
> 8229 107180 human p110d gene
8229 8479  utr
30002 30057 utr
69724 69826 utr
88681 88851 exon
93798 94024 exon
94106 94333 exon
94697 94874 exon
95216 95363 exon
95794 95881 exon
96951 97170 exon
98178 98272 exon
98369 98497 exon
98868 98916 exon
98999 99164 exon
99384 99503 exon
99701 99842 exon
100018 100115 exon
100232 100400 exon
100501 100611 exon
100785 100861 exon
101382 101547 exon
102226 102347 exon
102533 102676 exon
103061 103191 exon
105166 107180 utr
```

The sequence will be annotated in the appropriate position with the appropriate label. Vista Browser is designed to visualize multiple large-scale alignments. The browser's clean display makes it easy to identify regions of high conservation across multiple species. The browser is used to visualize pairwise and multiple whole genome alignments produced internally. The "peaks and valleys" graphs represent percent conservation between aligned sequences at a given coordinate on the base sequence. Multiple alignments that share the base sequence can be displayed simultaneously, one under another. The top and bottom percentage bounds are shown to the right of every row.

rVISTA (regulatory Vista) contains a list of 45 well-known transcription factor binding sites and upon selection of transcription factors of interest is combines transcription factor binding sites database search with a comparative sequence analysis.

2.11.4Transcription Element Search System (TESS)

TESS - **http://www.cbil.upenn.edu/tess/** is a web tool used for predicting transcription factor binding sites in DNA sequences. It identifies binding sites using positional weight matrices from the TRANSFAC, IMD, and the CBIL-GibbsMat databases. TRANSFAC® is one of the largest transcription factor databases specifying eukaryotic *cis*-acting regulatory DNA elements and *trans*-acting factors. It covers the whole range from yeast to human.

Submitting data is very self-explanatory. A web-link to the results is retuned via e-mail. The results are displayed in table format, containing the positions and orientation of the potential transcription factors. Links on each transcription factor supply more detailed information about the specifics of that factor, for example if it is species- or cell-type specific.

3 Investigation of a potential correlation between p110δ protein and mRNA levels

3.1 Introduction

Our investigation into the regulation of p110δ PI3K gene expression has revealed a complex regulatory system involving three untranslated exons and implies the existence of two p110δ promoters which are differentially used in several different cell types. The details of these regulatory regions are discussed in chapters 5, 6 and 7.

Like p110γ, p110δ has a restricted tissue distribution, with the highest levels seen in leukocytes (8). As discussed in the introduction, regulation of protein synthesis can occur at many different levels. We first used SDS-PAGE gel electrophoresis with immunoblotting to determine the level of mouse and human p110δ protein expression in different cell lines. Real time PCR analysis was used to analyse mRNA expression levels of the mouse and human p110δ gene. Comparative analysis was used to determine if there is a correlation between the RNA and protein levels of the p110δ gene.

Real time PCR is a more accurate method of determining the actual RNA levels of a particular gene than traditional PCR. Traditional PCR assays detect RNA/DNA levels at the end point of the reaction whereas real time PCR reactions analyse amplification during the exponential phase of the reaction.

3.2 Results

3.2.1 p110δ protein expression and mRNA levels in mouse and human cell lines

Since 1997 when the p110δ protein was first discovered its leukocyte-specific expression has been of great interest, particularly in the light of recent finding that p110δ plays a primary role in the immune responses. Using a refined strategy for the detection of p110δ protein in total cell lysates, we have now been able to detect p110δ in all cell types investigated, albeit at often very low levels. A table of cell lines, used in this and subsequent chapters, the organism from which they are derived and their corresponding cell type is shown in Table 2-1.

3.2.1.1 p110δ protein expression in mouse and human cell lines

Western blot analysis was performed on both mouse and human cell lines using either GAPDH or β-actin as a protein loading control. Due to a combination of reasons, one of which included the low amount of protein loaded per lane, early western blots for p110δ indicated that the p110δ protein was not expressed in non-leukocyte cell lines (an example is shown in Figure 3-1).

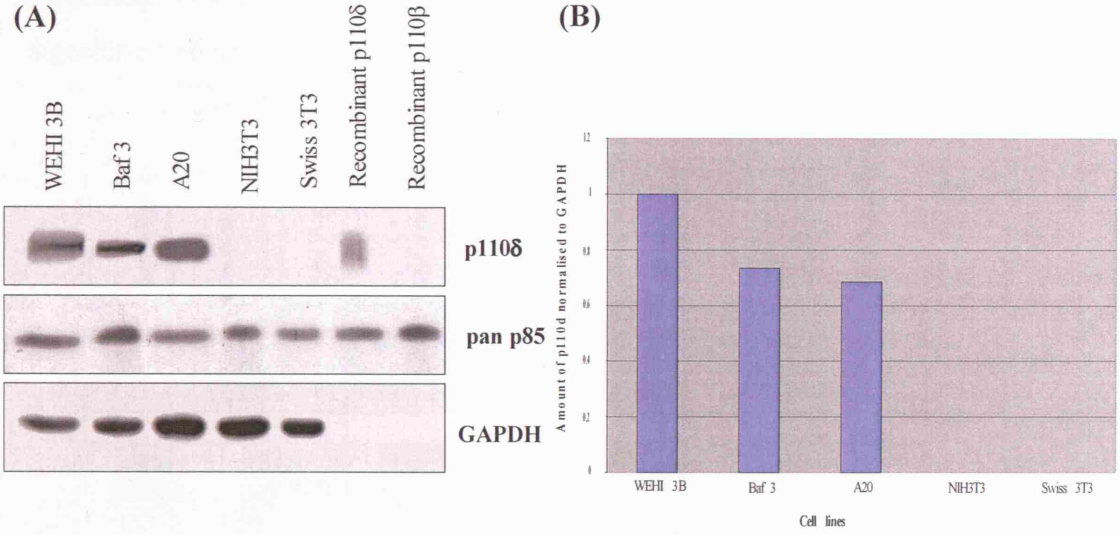

Figure 3-1 Differential protein expression between murine leukocyte cell lines (WEHI 3B, Baf 3 and A20) and non-leukocyte cell lines (NIH3T3 and Swiss 3T3). **(A)** 50 μg of total cell lysate from the indicated cell lines was resolved on SDS-PAGE gels and analysed for the expression of p110δ and p85 by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for GAPDH was used to assess protein loading. Recombinant p110δ and p110β were used as positive and negative protein loading controls for the p110δ antibody used for western blotting. **(B)** A graph of the p110δ signal normalise to the GAPDH signal using a densitometer to determine the strength of each of the bands. n= >5

The first evidence that p110δ protein was present in cell types other than leukocytes was when larger amounts of PI3K isolated from cell extracts were loaded on gels. This was achieved by immunoprecipitating p110δ first or by affinity purification of the cells Class IA PI3K. The latter was achieved by the use of PDGF receptor beads (platelet derived growth factor receptor), which allow the purification of the regulatory as well as the catalytic subunit of Class IA PI3Ks, and enrich for these proteins from total cell lysates. These peptides mimic the tyrosine-phosphorylated PDGF receptor and are coupled, via a spacer, to agarose beads. The beads are incubated with cell lysate. The SH2-domains of all Class IA regulatory subunits bind to the pYxxM sequence, allowing these proteins to be isolated using a pull down method, described in more detail in section 2.10.4. Baf 3 and WEH231 leukocytes and NIH3T3 fibroblasts, were cultured in their normal growth medium until 60% confluent followed by cell lysis, as described in section 2.10. The cell lysate from each cell line was divided equally between three eppendorf tubes; one was run on a gel as a total cell lysate control, one was used for p110δ immunoprecipitation and the third was incubated with PDGF receptor beads (Figure 3-2). These experiments identified the p110δ protein to be present in NIH 3T3 cells, albeit at a very low level. This prompted the need to refine the strategy to identify the p110δ protein in low expressing cell lines (strategy executed by other members of the Cancer Cell Signalling Group).

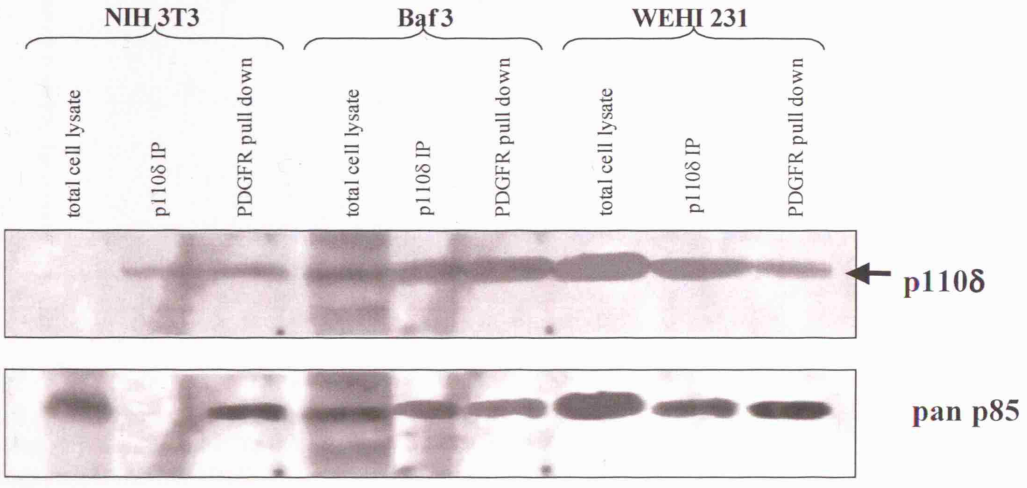

Figure 3-2 Evidence that mouse fibroblast cell lines express very low levels of p110δ. 70 μg of total cell lysate from the indicated cell lines was run in lanes 1, 4 and 7 and resolved on an SDS-PAGE gel. 1 mg of protein was used in each IP and PDGFR pull-down experiment and the resulting fraction resolved on an SDS-PAGE gel. The rat monoclonal p110δ antibody was used for the immunoprecipitation. The gel was analysed for the expression of p110δ and p85 by immunoblotting using a rabbit polyclonal antibody to p110δ. An arrow indicates the position of recombinant p110δ control. n = 1

A larger number of non-leukocyte cell lines were then analysed for p110δ protein expression and low levels were identified in the total cell lysate of all of the samples analysed. The screening of a large number of different cell lines confirmed the global expression of the p110δ protein in both mouse and human cells but at greatly differing levels (Figure 3-3).

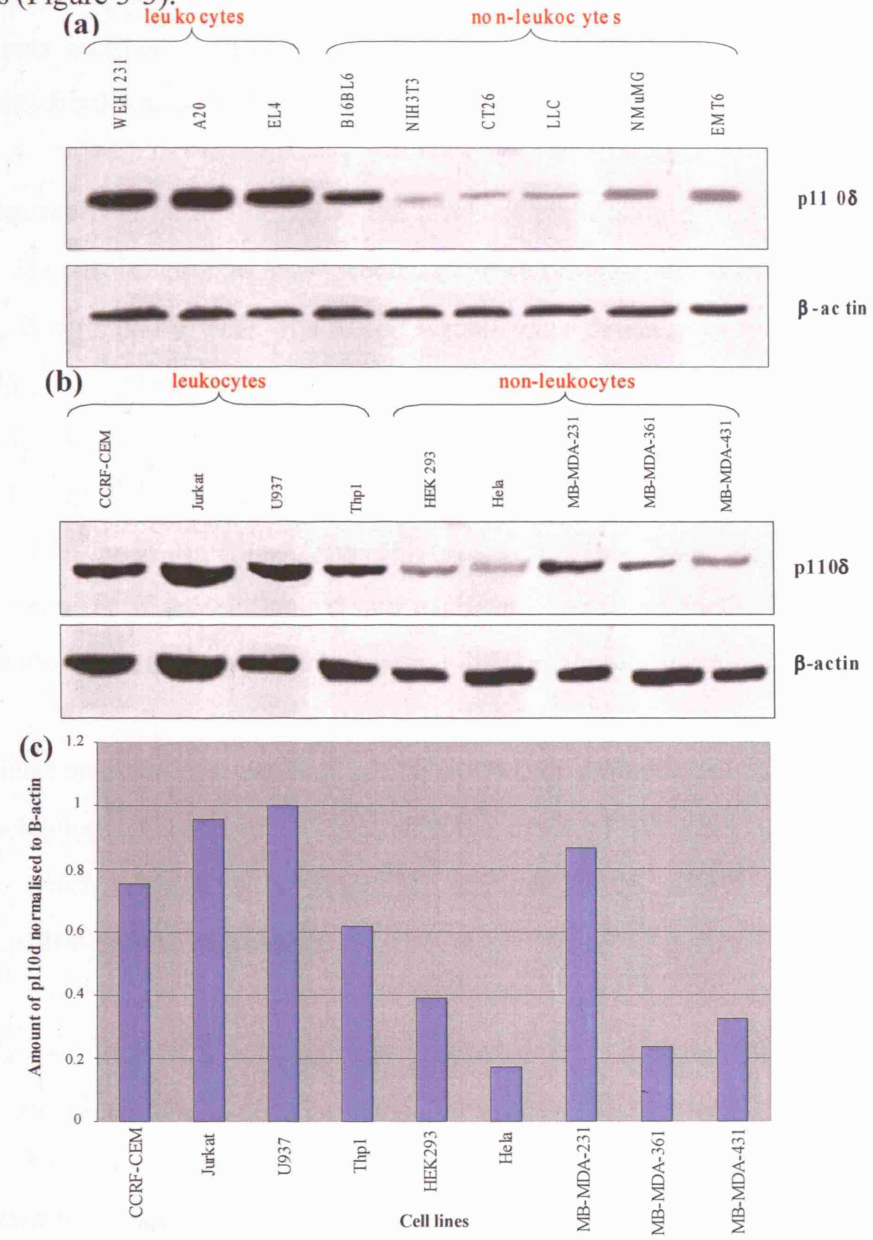

Figure 3-3 Differential protein expression between mouse and human leukocyte cell lines and non-leukocyte cell lines **(a)** A western blot of mouse cells. Source of cells: WEHI231, A20, EL4 (leukocytes), B16BL6, NIH3T3, CT26, LLC, NmuMG, and EMT6 (non-leukocytes) **(b)** A western blot of human cells. Source of cells: CCRF-CEM, Jurkat, U937, Thp1 (leukocytes), HEK, Hela, MB-MDA-231, MB-MDA-361 and MB-MDA-431 (non-leukocytes). 100 μg of total cell lysate from the indicated cell lines was resolved on SDS-PAGE gels and analysed for the expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for β-actin was used to assess protein loading. **(c)** A graph of the p110δ signal normalise to the β-actin signal, from the human cell line western blot, using a densitometer to determine the strength of each of the bands. n = 3

3.2.1.2 p110δ mRNA levels in mouse and human cell lines

The tissue distribution of p110δ had previously been investigated by northern blot analysis of human and mouse tissues (8,188). There was evidence of strong p110δ signals in tissues rich in leukocytes such as the spleen, thymus and peripheral blood leukocytes. Less obvious p110δ signal was recognizable in several other tissues but this was ascribed to possible blood cell contamination present in the crude tissue extracts from which the mRNA was isolated.

Vanhaesebroeck *et al.* looked at the levels of expression of the p110δ protein in cell lines (8). p110δ protein was present in both primary and transformed white blood cells, B cells and T cells. No p110δ signals were detected in fourteen non-leukocyte cell lines, including fibroblasts. We also showed high levels of p110δ protein in B and T cells (WEHI 231 and A20 Figure 3-3a) and in white blood cells (U937 Figure 3-3b). However we were able to detect low levels of p110δ protein in non-leukocyte cells lines such as fibroblasts (Figure 3-3). This was attributed to a higher concentration of protein loaded into each well, improved methods of signal detection and a more sensitive preparation method thus preventing protein degradation.

We next analysed the levels of p110δ mRNA in different cell lines. Recent advances in technology allowed us to determine the exact p110δ RNA levels in different cell lines, which enabled us to determine whether p110δ mRNA expression correlates with p110δ protein expression.

p110δ real time PCR primers were purchased from Applied Biosystems. The mouse forward primer bound to a sequence in exon 1 (incorporating the ATG translation start site). The reverse primer was complementary to exon 2. The internal primer spanned the junction between exon 1 and exon 2, thus increasing the specificity of the amplification. The human p110δ forward and reverse primers were complementary to sequences in exons 8 and 9 respectively with the internal primer spanning the exon-exon junction.

GAPDH was used as a reference gene to which the sample gene was normalized. p110α, thought to be ubiquitously expressed, was used as an internal control for the initial efficiency experiment (Figure 3-4).

The fluorescence is directly proportional to the amount of mRNA/cDNA that was present in the cell at the time of RNA extraction. The greater the mRNA expression, the sooner the PCR amplification occurs and the steeper the exponential phase. Knowing these facts, the differential p110δ mRNA expression between leukocytes and non-leukocytes was quite evident from the initial efficiency experiments (Figure 3-4).

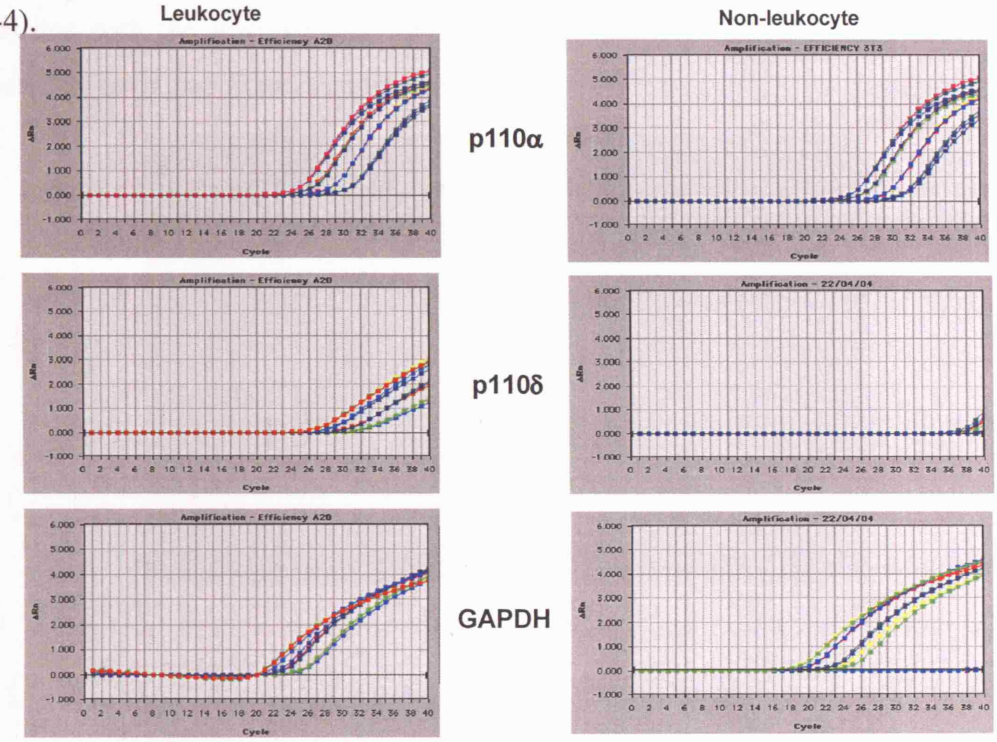

Figure 3-4 p110δ mRNA is expressed at much lower levels in non-leukocyte cell lines than in leukocyte cell lines. The graphs show the results from an efficiency experiment using the mouse A20 (leukocyte) cell line and the mouse NIH3T3 (non-leukocyte) cell line with real time PCR primers for mouse p110δ, p110α and GAPDH. The multiple lines on each graph correspond to different dilutions of cDNA. The X-axis represents the number of PCR cycles while the Y-axis records the level of fluorescence.

There are two methods of quantification namely the standard curve method, this uses a series of dilutions of standard cDNA with which to compare the target cDNA. Secondly there is the comparative C_T method ($\Delta\Delta C_T$), this uses a series of arithmetric formulas to calculate the relative amount of target cDNA. For the latter method to be valid, the efficiency of the target amplification and the efficiency of the reference amplification must be approximately equal.

Relative efficiency plots were prepared, from standard curve experiments, for all primer sets to ensure the $\Delta\Delta C_T$ relative quantification method could be applied to the genes of interest with statistical significance (Figure 3-5). For this to be applied, the absolute value of the slope of the logarithm of the input amount (expressed in ng) vs. ΔC_T should be < 0.1. Standard curves are explained in greater detail in section 2.8.2.1.

The results of the relative efficiency plots confirmed the efficiency of the mouse (slope = 0.0988) and human (slope = 0.1001) p110δ and β-actin real time PCR primers therefore $\Delta\Delta C_T$ relative quantification was used for the RNA analysis. The cell line with the lowest expression level of p110δ was delegated as the calibrator, therefore having an expression level of 1.

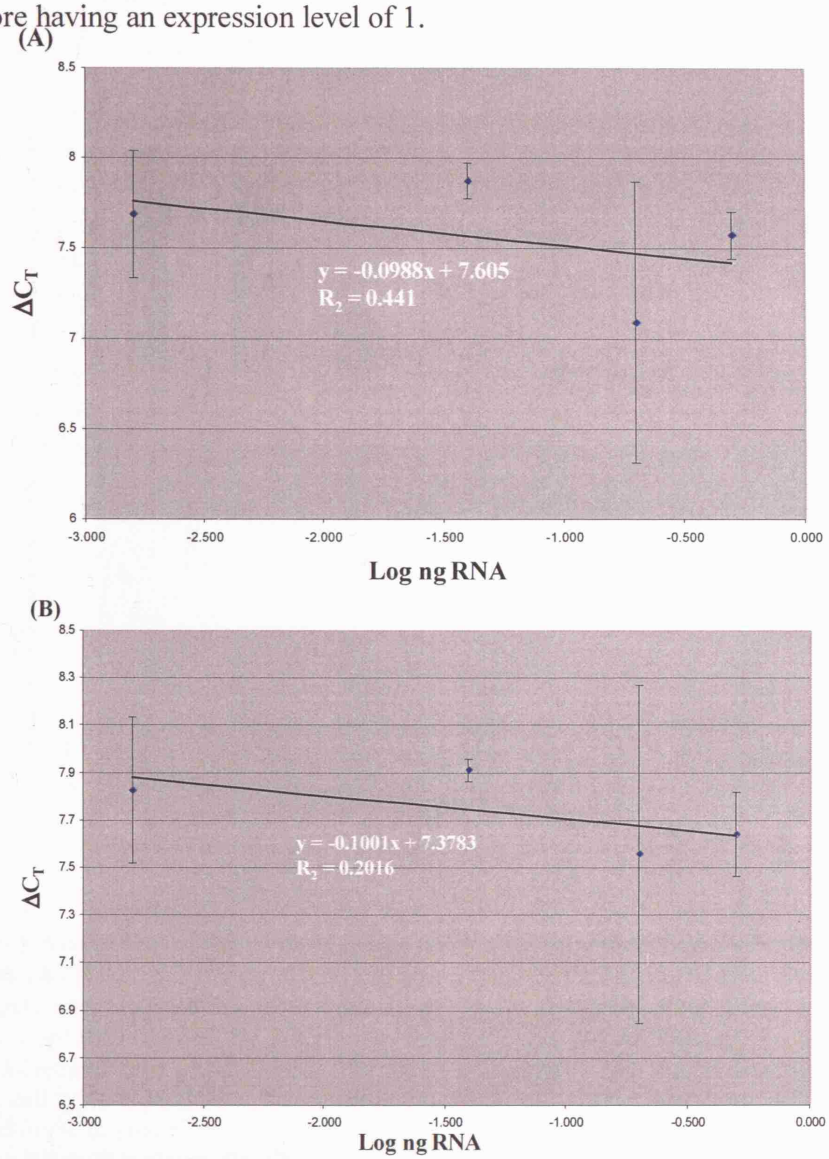

Figure 3-5 Relative efficiency plots. (A) Relative efficiency plot of mouse p110δ and mouse β-actin. (B) Relative efficiency plot of human p110δ and human β-actin.

Seven mouse and seven human cell lines were analysed for p110δ protein and mRNA expression simultaneously. In the mouse, a clear correlation between p110δ protein and mRNA levels was evident (Figure 3-7). The intermediate level of p110δ protein expression seen in the B16-BL6 melanoma cell line was also seen in other cell lines of a similar type (B16-F10 and melan-b, data not shown). In the human cell lines, such a correlation was not as clear-cut (Figure 3-7).

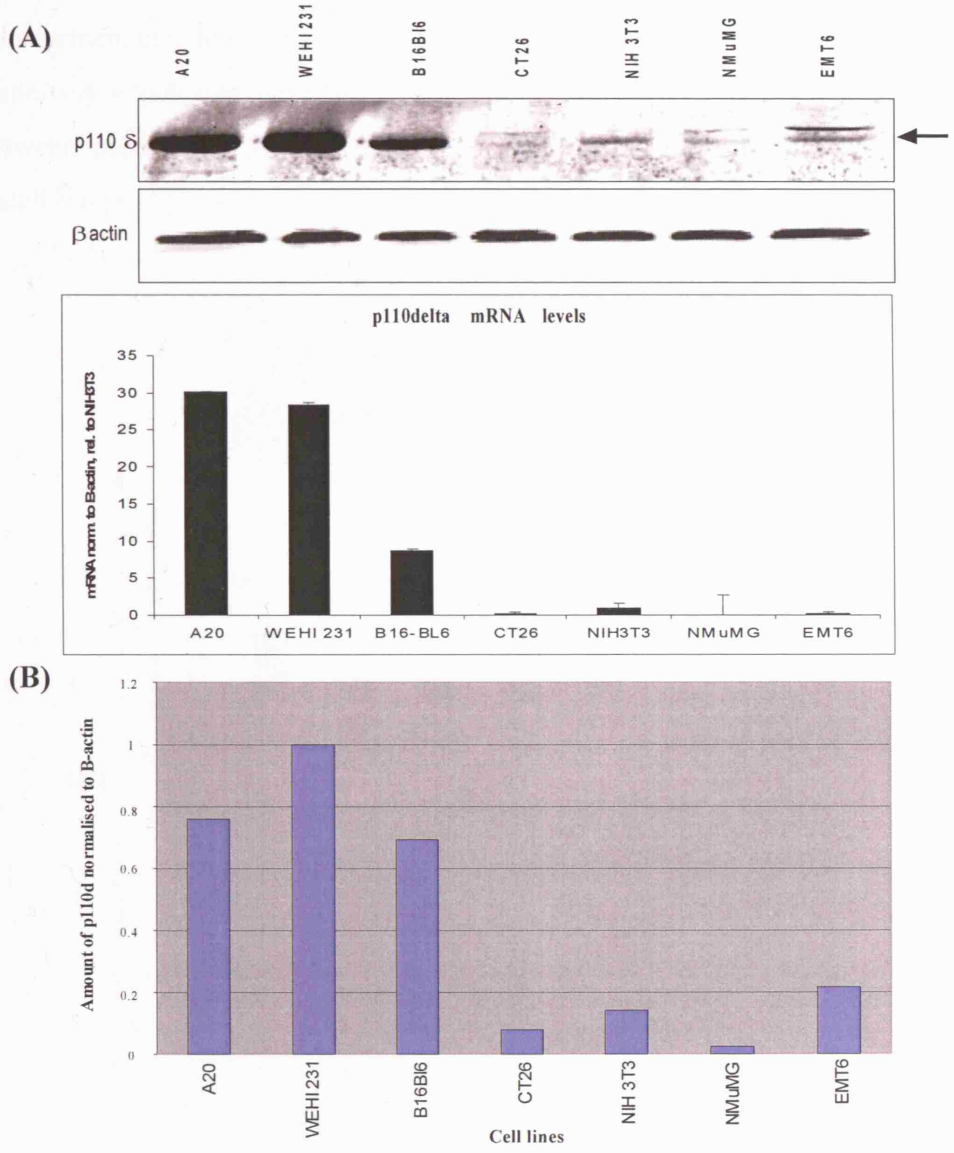

Figure 3-6 Assessment of the levels of mouse p110δ protein and mRNA expression levels in mouse cell lines. (A) *Top panel,* 100 µg of total cell lysate from the indicated cell lines was resolved on SDS-PAGE gels and analysed for expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for β-actin was used to assess protein loading. An arrow indicates the position of recombinant p110δ control. The nature of the higher MW signal detected by p110δ antisera in some cell types is unclear at the moment. *Bottom panel,* mRNA was extracted from the cell lines indicated in the top panel, reverse transcribed to cDNA and equal amounts of cDNA used in a 60 cycle Real time RT-PCR reaction. Results were normalized to β-actin as a control. **(B)** A graph of the p110δ protein signal normalised to the β-actin protein signal using a densitometer to determine the strength of each of the bands. n = 4

It is not clear why the human p110δ protein levels do not correlate with the mRNA levels quite as well as they do in the mouse cell lines studied. This may well be due to the small sample size tested. If a larger number of human cell lines had been tested for p110δ protein and mRNA expression a better correlation may have become apparent. We may have been lucky with the choice of mouse cell lines. All of the cell lines tested for a correlation between p110δ protein expression and mRNA levels were transformed cell lines therefore, at one time, had had their regulation disrupted in some way which may have had a knock on effect and effected the *in vivo* correlation between protein and mRNA expression. This would have affected all the cell lines tested but possibly some more than others.

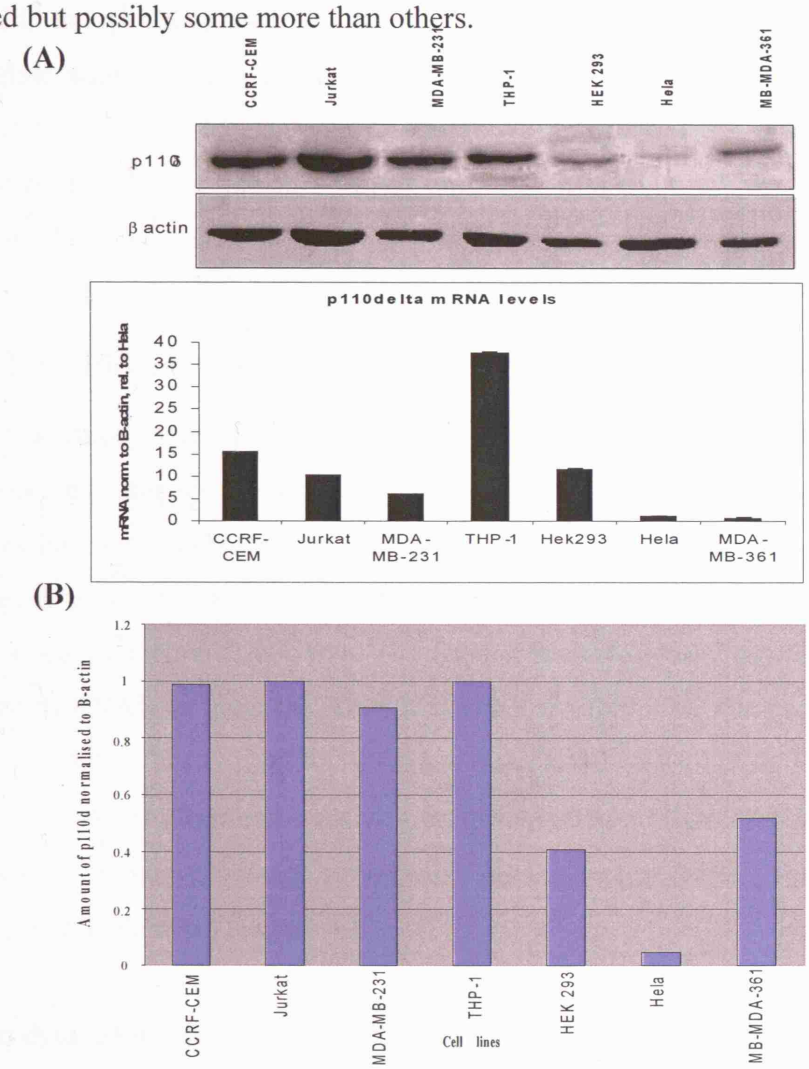

Figure 3-7 Assessment of the levels of human p110δ protein and mRNA expression levels in cell lines. (A) *Top panel,* 100 µg of total cell lysate from the indicated cell lines was resolved on SDS-PAGE gels and analysed for expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for β-actin was used to assess protein loading. *Bottom panel,* mRNA was extracted from the cell lines indicated in the top panel, reverse transcribed to cDNA and equal amounts of cDNA used in a 60 cycle Real time RT PCR reaction. Results were normalized to β-actin as a control. **(B)** A graph of the p110δ protein signal normalised to the β-actin protein signal using a densitometer to determine the strength of each of the bands. n = 2

Taking into account past and recent evidence we would expect the levels of p110δ protein to be low in the HEK 293 cell line. However the β-actin signal in the HEK 293 sample is much lower than the signal seen in the leukocyte samples. This implies that less protein was loaded onto the gel. If this was the case, and a similar concentration of protein was loaded onto the gel, the signal would have been stronger and would therefore correlate more closely with the RNA levels.

However a weaker β-actin signal seen in the MDA-MB-361 lane suggests that there was a similar problem with the loading. If a similar concentration of protein had been loaded, the p110δ protein signal would have been stronger and would then not correlate with the RNA levels. It is possible that some of the more spurious results may have been due to experimental error. Working with RNA along with very small volumes is not easy due to possible problems with degradation. Therefore a repeat experiment is required.

3.2.2 p110δ protein expression and mRNA levels in mouse organs

In an attempt to resolve the possible issue of disrupted regulatory mechanisms in transformed cells mouse organs were analysed for a correlation of p110δ protein expression and mRNA levels. Two 10-week-old C57Bl6/J wild-type mice were perfused with PBS to try and remove as may of the red blood cells as possible (described in section 2.4.3.4.1). The organs were removed; one set of organs was placed in RNA*later*® and the RNA was extracted following the protocol described in section 2.4.4.3. The second set of organs was placed on ice before being homogenized and lysed for p110δ protein analysis, as described in section 2.10.2. The results from both sets of analysis were inconclusive because of possible contamination of the tissue with leukocytes (Figure 3-8).

Leukocyte contamination was analysed by immunoblotting for CD45. The CD45 antigen is a tyrosine phosphatase, also known as the leukocyte common antigen (LCA). CD45 is present on all human cells of hematopoietic origin except platelets. The antibody used recognizes a common structure of the 5 different isoforms of CD45.

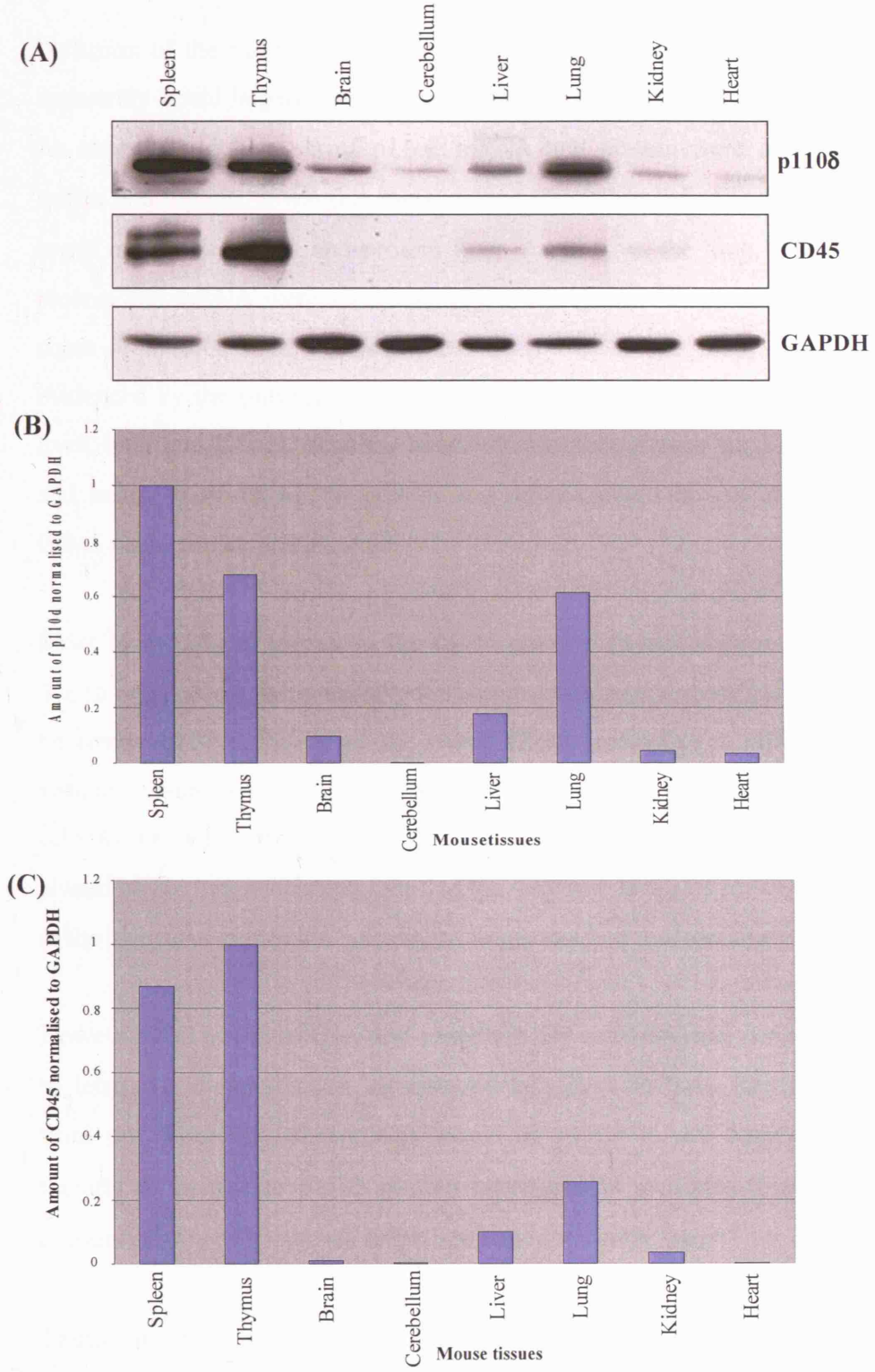

Figure 3-8 p110δ protein expression in mouse tissues. 100 μg of total cell lysate from the indicated mouse organs was resolved on SDS-PAGE gels and analysed for the expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Leukocyte contamination was analysed by immunoblotting for CD45. Blotting for GAPDH was used to assess protein loading. **(B)** A graph of the p110δ protein signal normalised to the GAPDH protein signal using a densitometer to determine the strength of each of the bands. **(C)** A graph of the CD45 protein signal normalised to the GAPDH protein signal using a densitometer to determine the strength of each of the bands. n = 2

Perfusion of the mice to flush out the circulating leukocytes was carried out, as this apparently could largely overcome the problem of leukocyte contamination of tissues. As expected, high levels of p110δ mRNA and protein were identified in both the spleen and thymus, given that these organs mainly consist of leukocytes. Intermediate levels of p110δ mRNA and protein were identified in the lung. Low levels pf p110δ protein and mRNA were also identified in the kidney, heart, brain and cerebellum. In some of these tissues, this was most likely due to the presence of leukocytes, as evidenced by the presence of expression of CD45 (a leukocyte marker) in lysates of liver, lung and kidney samples. However, in other tissues such as brain, cerebellum and heart, in which p110δ protein and mRNA could be detected, the presence of CD45 could not be documented.

Most likely, p110δ signals in the CD45-positive tissues such as liver and lungs are due to presence of tissue resident leukocytes such as macrophages etc which could not be removed by perfusion of the mice. These macrophages differ in appearance in various tissues and are known by various names: Kupfer cells in the liver, reticular cells in the lymph nodes, spleen and bone marrow and alveolar macrophages in the alveoli of the lungs. It is possible that the very low levels of p110δ mRNA and protein in the heart and kidney are also due to tissue resident macrophages.

However, the p110δ mRNA and protein in the brain and the cerebellum was not due to leukocyte contamination, as assessed by CD45 analysis. Recent unpublished data from our laboratory shows high levels of p110δ in the brain and spinal column thought to be due to p110δ protein expression in neurones (Figure 3-9). This may account for the p110δ bands in the brain and the cerebellum.

Transgenic mice were created with a copy of the β-galactosidase gene inserted into the p110δ coding region before the STOP codon. Therefore β-galactosidase expression coincided with p110δ protein expression. The embryos were fixed and stained with X-gal. β-galactosidase turns X-gal into an insoluble blue dye. This allows the sites of beta galactosidase expression, and therefore p110δ expression, to be visualised in the mouse embryo (Figure 3-9).

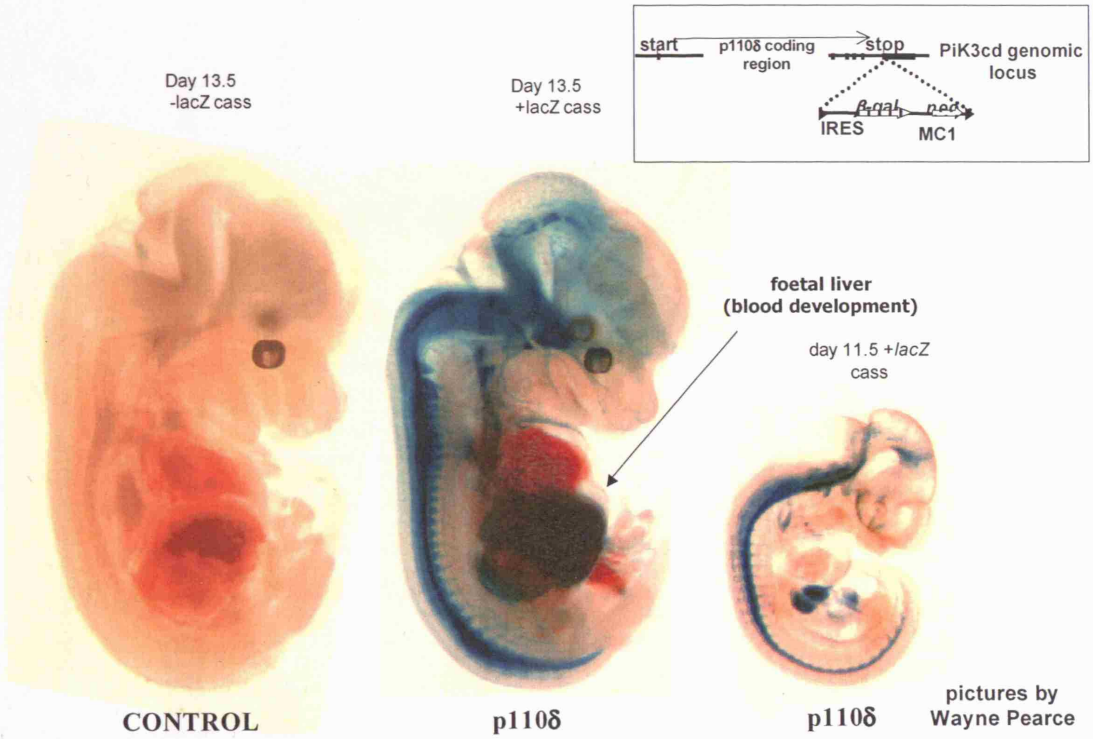

Figure 3-9 p110δ expression in a mouse embryo (pictures provided by Wayne Pearce, Ludwig Institute for Cancer Research, London). Transgenic mice were created with a copy of the β-galactosidase gene inserted into the p110δ coding region before the STOP codon. The inset picture is a schematic representation of the position of the β-galactosidase gene in relation to the p110δ coding region and STOP codon.

Blood development occurs in the foetal liver resulting in the strongest staining seen here (Figure 3-9). Strong X-Gal staining is also evident in the spinal column and in the brain; this is believed to be due to high levels of p110δ expression in neurones. The negative control is very clear, ruling out the possibility of any of the blue being as a result of background staining from endogenous β-gal activity.

The p110δ mRNA levels in total RNA extracted from different mouse organs was also analysed by real time PCR analysis, as described in section 2.8.2.2. The RNA levels correlated with the protein levels seen using SDS PAGE analysis and immunoblotting (Figure 3-10).

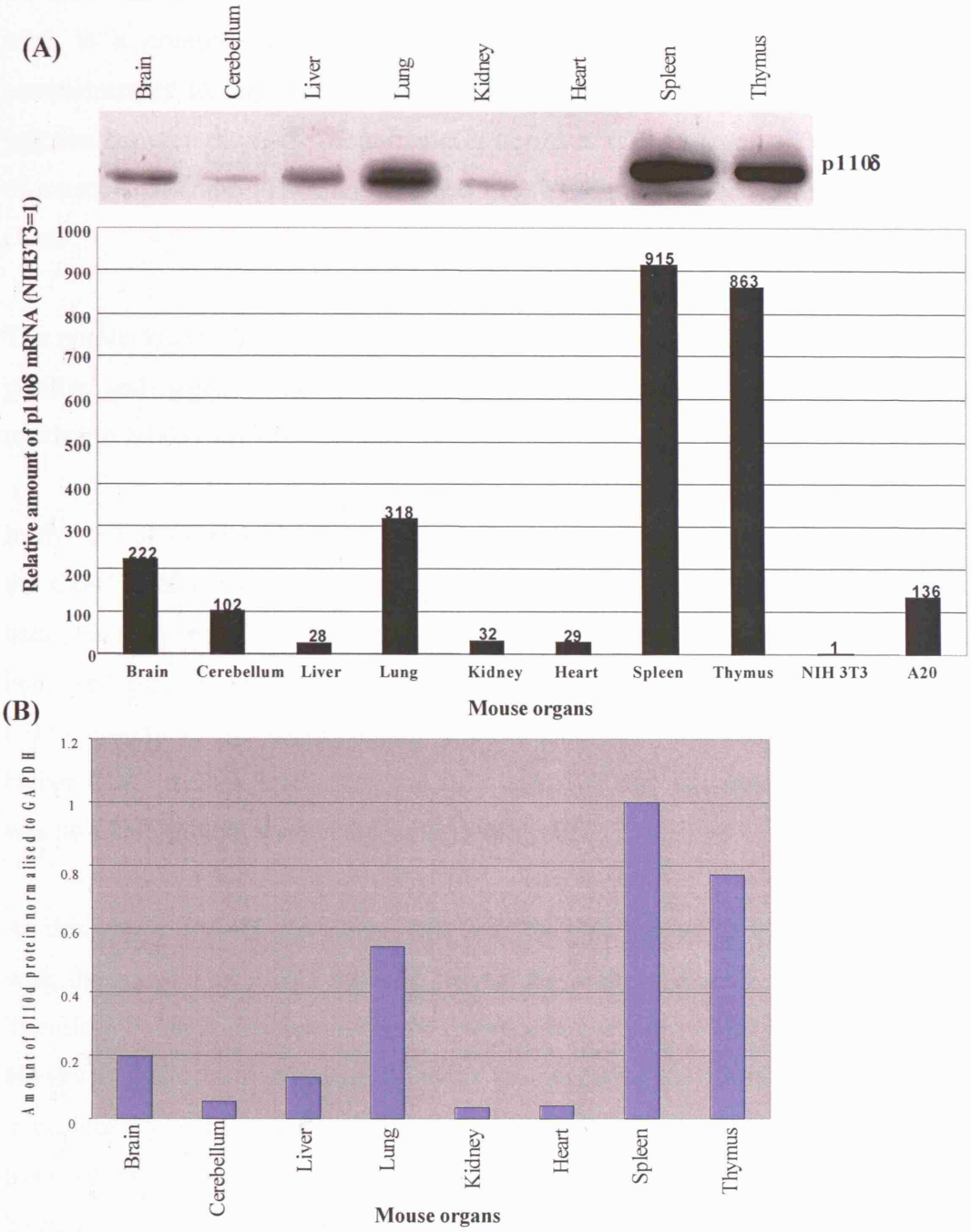

Figure 3-10 p110δ protein expression correlates with p110δ mRNA levels in eight mouse organs analysed. (A) *Top panel*: p110δ protein expression in eight different mouse organs analysed for the expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ (as seen in Figure 3-8). *Bottom panel*: p110δ real time PCR primers specific for p110δ cDNA were used to analyse the indicated mouse organs. 5 μg of total RNA was reverse transcribed to cDNA using an oligo dT primer. 1μl of the cDNA was then subjected to 60 rounds of PCR amplification. The level of amplification was recorded during the exponential phase of the PCR reaction. The relative levels of p110δ mRNA were calculated with the level of p110δ mRNA in NIH3T3 set as 1. **(B)** A graph of the p110δ protein signal normalised to the GAPDH protein signal using a densitometer to determine the strength of each of the signals. n = 2

To assess the level of leukocyte contamination real time PCR primers for CD45 were used in a control experiment using the same cDNA. The CD45 primers are complementary to sequences in exons 2 and 3 with the internal primer spanning the junction between the two. The different isoforms of CD45 arise from variable splicing of exons 4, 5, and 6. Therefore the real time PCR primers recognise all isoforms of CD45.

The non-leukocyte cell line, NIH3T3, and the leukocyte cell line, A20, were used as positive and negative controls, respectively. The expression levels were calculated relative to NIH3T3 (Figure 3-11).

In five out of the eight mouse organs analysed, liver, lung, kidney, spleen and thymus, the CD45 protein expression levels correlated with the CD45 mRNA levels. In the brain, cerebellum and the heart they did not. CD45 mRNA was present in the mouse brain and cerebellum, albeit at relatively low levels. Yet there was no evidence of CD45 protein in the corresponding protein analysis. The mouse heart contained higher CD45 mRNA levels than the A20 leukocyte cell line sample, but again there was no CD45 protein band in the western analysis.

As this was a one-off experiment it is possible that there may have been a problem with the transfer of CD45 protein, during the western blot, in some of the wells. Therefore before a conclusion can be drawn a repeat experiment must be undertaken. However, from these experiments it can be concluded that even after perfusing the mice, leukocyte contamination still persists. This is most likely due to the presence of tissue-resident cells (such as macrophages), which unlike cells in circulation, cannot be removed by the perfusion of mice.

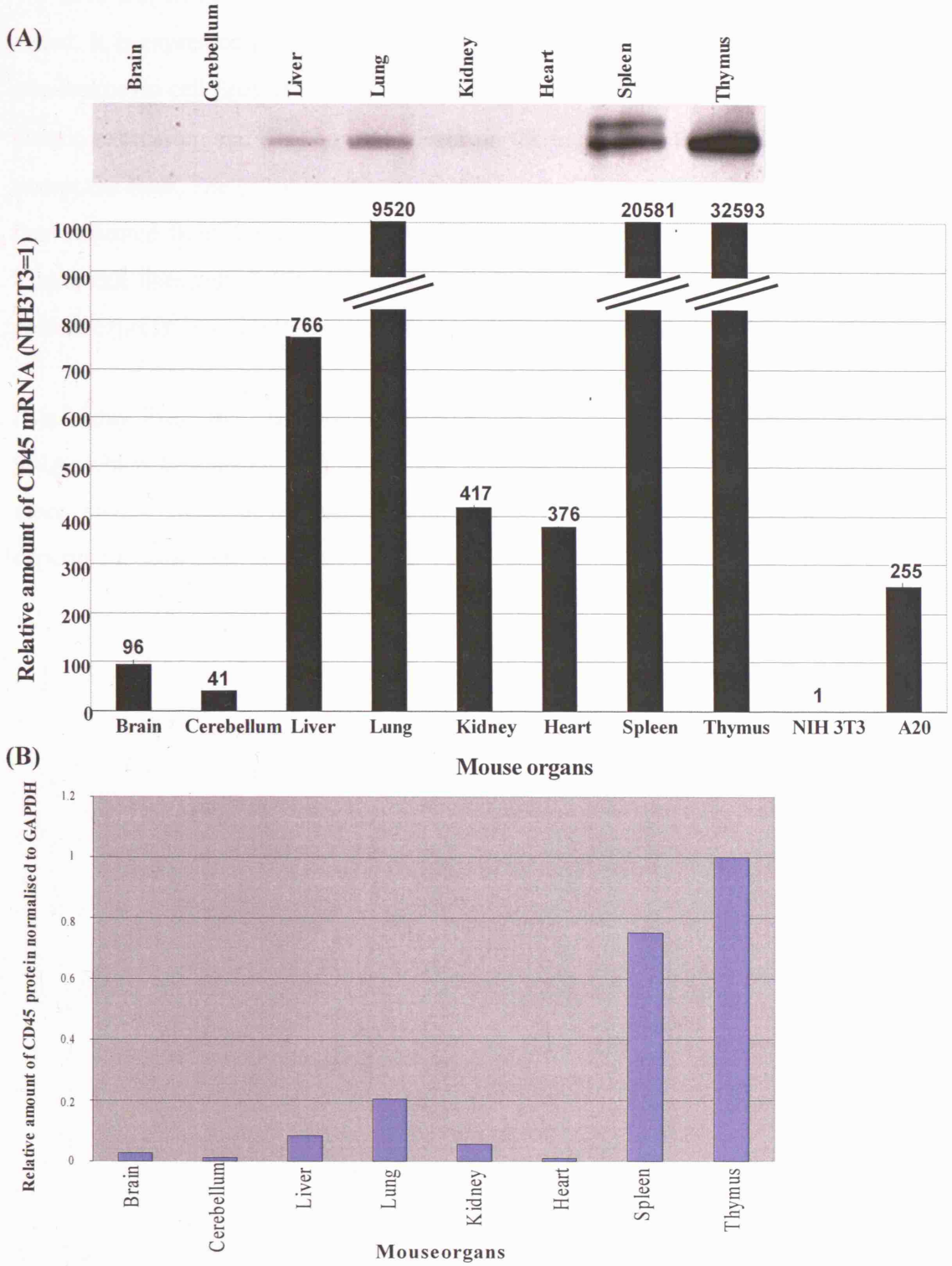

Figure 3-11 CD45 protein expression does not always correlate with CD45 mRNA levels in eight mouse organs analysed. (A) *Top panel*: CD45 protein expression in eight different mouse organs analysed for the expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ (as seen in Figure 3-8). *Bottom panel*: real time PCR primers specific for CD45 cDNA were used to analyse the indicated mouse organs. 5 μg of total RNA was reverse transcribed to cDNA using an oligo dT primer. 1 μl of the cDNA was then subjected to 60 rounds of PCR amplification. The level of amplification was recorded during the exponential phase of the PCR reaction. The relative levels were calculated with NIH3T3 levels set as 1. **(B)** A graph of the CD45 protein signal normalised to the GAPDH protein signal using a densitometer to determine the strength of each of the signals. n = 1

We have shown that p110δ protein is expressed in all human and mouse cell lines tested. It is expressed predominantly in cells of leukocyte origin but is also present in non-leukocyte cell lines at a much lower level. There is a correlation between p110δ protein expression and p110δ mRNA levels in the mouse cell lines and tissues tested, except the liver. The p110δ protein expression in the liver sample appears to be higher than expected from the real time PCR results. The correlation is less clear-cut in the human cell lines tested. There is a trend but not a direct correlation between p110δ protein expression and mRNA levels in the human cell lines tested.

It is highly likely that these slight discrepancies are due to technical issues in using RNA, which is unstable, and p110δ that appears to be a fairly unstable protein. A more sterile working environment, improved laboratory techniques and repeat experiments may solve these inconsistencies.

4 Expression of p110δ is not altered by acute cellular stimulation

4.1 Introduction

Throughout the life cycle of a cell different combinations and permutations of proteins are required for ongoing cellular activities and are needed to respond to numerous incoming signals. Many proteins are required all of the time, but for those that are not, the cell has various mechanisms to up and down-regulate activity when required. Some of those mechanisms include activating/deactivating a protein that is already in the cell, increasing/decreasing the rate of synthesis of the protein and targeting/not targeting the protein for degradation.

As described in section 1.9 very little is known about the regulation of PI3K. The mechanism of gene expression of the p110 subunits and their regulatory subunits also is unknown. We investigated whether p110δ protein expression can be altered in cells by 1) various stimuli, such as stress and antigen receptor activation and 2) during different phases of the cell cycle upon the global experimental modification of the chromatin structure.

Stimuli that were used:

Tumour necrosis factor-α (TNF-α) is elevated in acute inflammatory states. Virtually all human tissues express tumour necrosis factor receptor 1, TNFR1, the major signalling receptor for TNF-α. TNFR2, which binds both TNF-α and TNF-β, is chiefly expressed on immune cells. TNF-α exposure can result in activation of a caspase cascade leading to apoptosis, or it can have the opposite effect, activation of transcription factors that result in cell survival/proliferation. TNF-α also induces genes involved in acute and chronic inflammatory responses (189). Given the expression of p110δ in leukocytes, we wanted to explore if p110δ was amongst the TNF-responsive genes.

Osmotic stress occurs when the concentration of molecules in solution outside of the cell is different than that inside the cell. When this happens, water flows either into or

out of the cell by osmosis, thereby altering the intracellular environment. Hyper-osmotic stress causes water to diffuse out of the cell, resulting in cell shrinkage, which can lead to DNA and protein damage, cell cycle arrest, and ultimately cell death. Cells compensate or adapt to osmotic stress by activating an osmotic stress response pathway that is controlled by a gene called nuclear factor of activated T cells 5 (NFAT5)/tonicity enhancer binding protein (TonEBP). This NFAT5/TonEBP protein is the only known mammalian transcription factor that is activated by hyper-osmotic stress (190). Given the implication of PI3K in cell survival, we considered that cells might upregulate p110δ protein in order to protect themselves during stress responses.

UV radiation can induce an inflammatory response, sunburn. Acute UV radiation exposure induces the migration of inflammatory cells into the exposed area. TNF-α and a number of other cytokines play a key role in this response. Exposure of HaCaT (human epidermal) cells to UV irradiation lead to a dose-dependent induction of TNF-α mRNA levels, at certain time points (191). UV irradiation can induce acute effects in the skin, such as membrane damage resulting in the activation of the transcription factor nuclear factor kappa B (NFκb), activation of Src tyrosine kinases, and DNA damage repair proteins (192).

Protein degradation is a common mechanism used to regulate cellular pathways. Controlled degradation by the proteasome is a key regulatory step in the execution of many cellular processes, including cell proliferation and apoptosis. Proteasome inhibitors have been shown to stabilise a number of key proteins such as tumour suppressor protein p53, cell cycle regulator p27 and pro-apoptotic proteins Bid and Bax (193).

Targeting the ubiquitin-proteasome pathway has recently emerged as a promising approach to the treatment of several diseases, including cancer and inflammatory disease. Initial efforts have focused on the proteasome, a large multisubunit protease complex required for regulated protein turnover. In clinical trials, the proteasome inhibitor PS-341 has shown promising anti-tumour activity in multiple cancers (194).

We considered that p110δ might be absent from some cells due to constitutive degradation of the p110δ protein, and therefore aimed to interfere with the cellular machinery for protein degradation.

Demethylation and inhibition of deacetylation: DNA methylation and histone acetylation are important epigenetic mechanisms that control gene expression by dictating transitions between transcriptionally active or transcriptionally silent chromatin states (195,196). Certain chemical agents, such as 5'-azacytidine (5AC) or Trichostatin A (TSA) can be used to interfere with these epigenetic mechanisms. 5AC causes DNA demethylation or hemi-demethylation, which creates a more relaxed chromatin structure allowing transcription factors to bind to DNA, while TSA inhibits class I and II histone deacetylases. The resulting histone hyperacetylation leads to chromatin relaxation and modulation of gene expression. We considered that if p110δ protein expression is regulated by such epigenetic mechanisms we might observe a difference in p110δ protein expression upon interference with such global mechanisms of gene regulation.

Anti-IgM: Signalling through the B-cell antigen receptor (BCR) is a key determinant in the regulation of B cells and can trigger B lymphocyte activation, proliferation, or apoptosis. Engagement of the BCR by Ag or anti-receptor Abs can lead to multiple cellular responses. In B cells, the PH domain-containing kinase Akt is activated in a PI3K-dependent manner (197). Array data characterising gene expression in B-cell malignancies reported a increase in p110δ expression after anti-IgM antibody stimulation for 24 hours (198). Given the very strong phenotype of p110δ inactivation in mouse B cells (152) we considered that B cells might regulate p110δ protein expression in a dynamic manner.

Mineralocorticoid: Recent reports presented evidence for the induction of the p110δ protein in rat cardiac fibroblasts after stimulation with aldosterone, a mineralocorticoid (199). Overexpression of mineralocorticoid receptors (MRs), responsible for the genomic effects of aldosterone, causes cardiomyopathy in mice. Inhibition of p110δ stopped the hypertrophic and profibrotic effects.

This chapter describes the different experiments done in an attempt to increase p110δ expression in NIH3T3 and L929 mouse cell lines and primary B cells using the different stimuli mentioned above.

4.2 Results

For our studies to up-regulate p110δ protein expression we used mainly NIH3T3 mouse fibroblast cells which normally only express their protein at a very low level, as documented in chapter 3.

For each set of stimulations one million NIH3T3 cells were added to 10 cm round tissue-culture dishes the night before stimulation and allowed to adhere. Experiments were repeated numerous times to determine the optimal response time to stimulation and the amount or concentration of the stimuli. The cells were stimulated in staggered intervals to ensue that cell lysis occurred simultaneously. Two plates remained unstimulated, one of which was lysed at the beginning of the experiment and one at the end to assess the impact of confluency of the cell culture.

4.2.1 p110δ protein levels are not altered by TNF-α

Stimulations occurred over a 24 h period before the cells were lysed and 100 μg of total cell lysate was resolved on an SDS-PAGE gel and analysed for expression of p110δ by immunoblotting (Figure 4-1). TNFα did not appear to have any effect on p110δ expression. The slight increase in p110δ protein in a few of the lanes was due to loading differences (controlled for by assessment of GAPDH expression). 1) It would have been advisable to have carried out an immunoblot for a TNF-α induced response as a positive control to see if the TNF stimulation worked. 2) This could have been followed up by looking at genes known to be induced by TNF-α such as IL6, IL-1β and MAP3K (200).

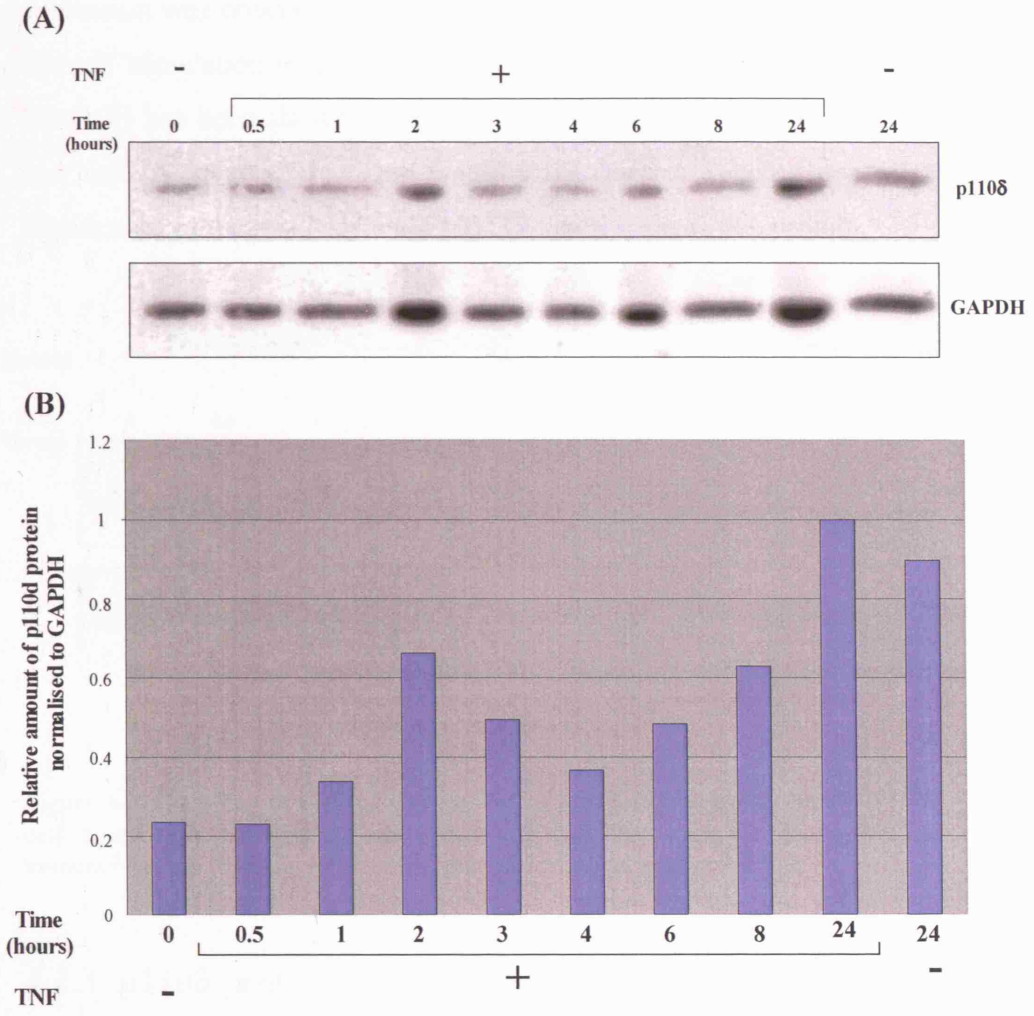

Figure 4-1 The effect of TNF-α stimulation on p110δ expression in NIH3T3 cells. 100 μg of total cell lysate was resolved on an SDS-PAGE gel and analysed for expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for GAPDH was used to assess protein loading. **(B)** A graph of the p110δ protein signal normalised to the GAPDH protein signal using a densitometer to determine the strength of each of the signals. n = 1

4.2.2 p110δ protein levels are not altered by osmotic stress

An experiment identical to the TNFα stimulation was set up using 4 M sorbitol instead of TNFα. The sorbitol exerts hyper-osmotic stress on the cell that causes water to exit the cell and the cell to shrink. Again, as controls, two plates remained unstimulated, one of which was lysed at the beginning of the experiment and one at the end.

The cells were lysed and 100 μg of total cell lysate was resolved on an SDS-PAGE gel and analysed for expression of p110δ by immunoblotting. No increase in p110δ

expression was observed over 24 h (Figure 4-2). However an appropriate control for the cell stimulation was not used. Hypotonic stress applied to human epithelial cells (HaCaT) has been shown to induced upregulation of E-cadherin at both the protein and mRNA level (201). This would have been a suitable control to ensure the significance of the results, should NIH3T3 cells express this protein.

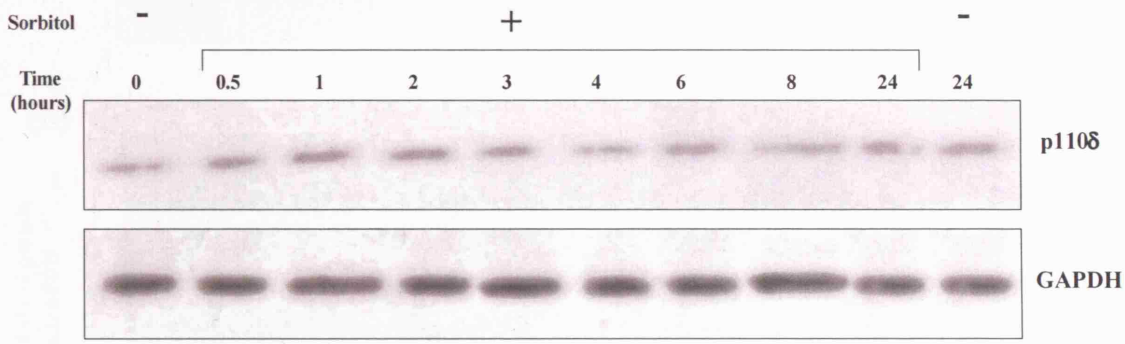

Figure 4-2 The effect of sorbitol stimulation on p110δ expression in NIH3T3 cells. 100 μg of total cell lysate was resolved on an SDS-PAGE gel and analysed for expression of p110δ by immunoblotting. Blotting for GAPDH was used to assess protein loading.

4.2.3 p110δ protein levels are not altered by UV radiation

The experiment was executed over a shorter time course than that of the TNF and sorbitol experiments due to the toxicity of UV radiation over a 24 h period.

NIH3T3 cells exposed to UV light at a wavelength of 524 nm over a period of between 5 min –30 min showed no increase in p110δ protein expression. Previous optimisation experiments were done to calculate the correct length of exposure of the cells to the radiation. It would have been desirable to have carried out an immunoblot for TNF-α or IL-1β, both induced by UV stimulation (191), as a positive control.

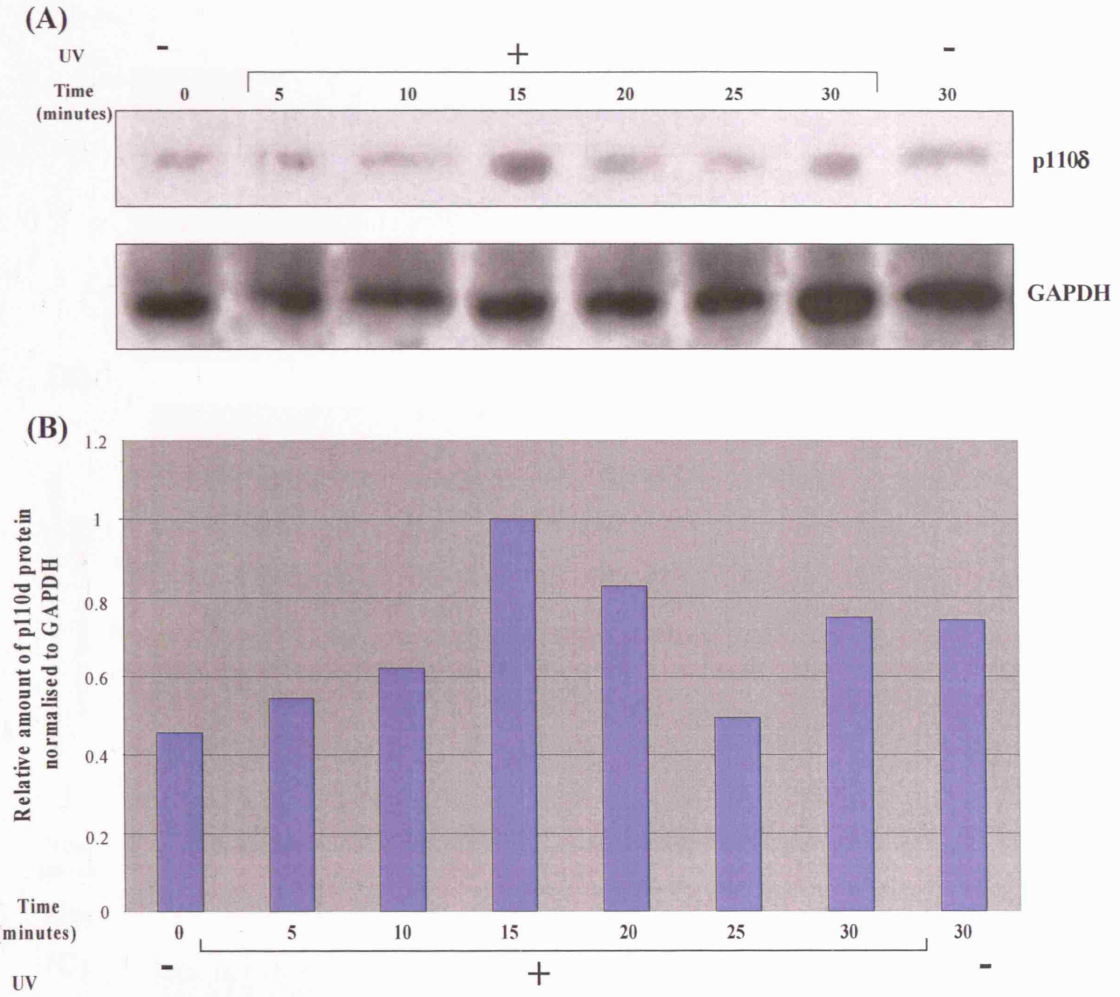

Figure 4-3 The effect of UV radiation on p110δ expression in NIH3T3 cells. 100 μg of total cell lysate was resolved on an SDS-PAGE gel and analysed for expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for GAPDH was used to assess protein loading. **(B)** A graph of the p110δ protein signal normalised to the GAPDH protein signal using a densitometer to determine the strength of each of the signals. n = 1

4.2.4 p110δ protein levels are not altered by proteasome inhibition

Using existing knowledge of the optimum experimental conditions for PS-341 (194) combined with in-lab experience of PI3K and proteasome inhibition, an experiment was devised to determine whether p110δ protein expression is affected by proteasome inhibition. 100 μM PS-341 was added to each of the plates over a period of 8 h. One plate was left unstimulated as a control (Figure 4-4).

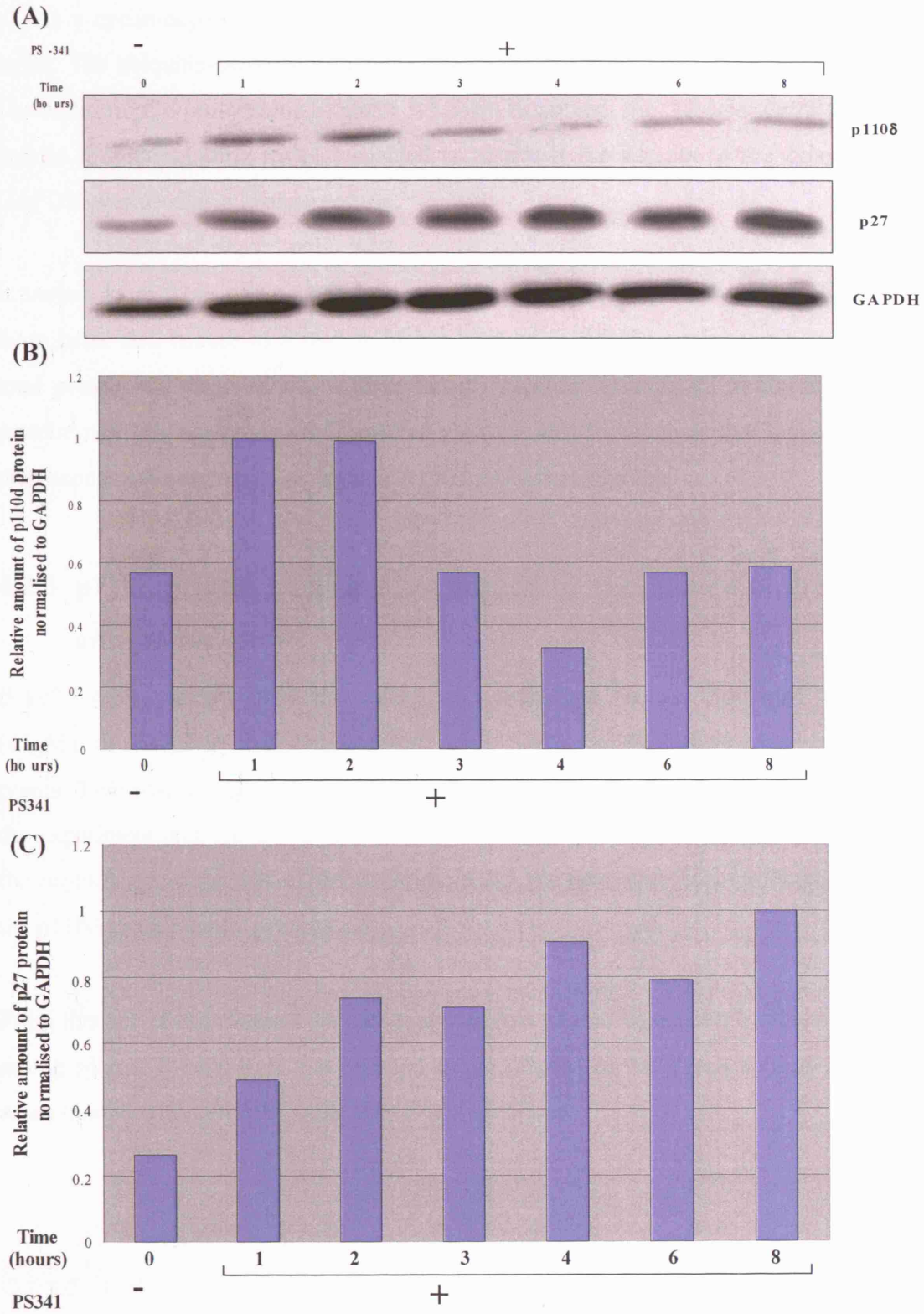

Figure 4-4 The effect of proteasome inhibition on p110δ expression in NIH3T3 cells. 100 μg of total cell lysate was resolved on an SDS-PAGE gel and analysed for expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for GAPDH was used to assess protein loading. p27 was used as a positive control to assess the success of the stimulation. **(B)** A graph of the p110δ protein signal normalised to the GAPDH protein signal using a densitometer to determine the strength of each of the signals. **(C)** A graph of the p27 protein signal normalised to the GAPDH protein signal using a densitometer to determine the strength of each of the signals. n = 1

p27 is a cyclin-dependent kinase inhibitor which upon activation leads to cell cycle arrest. The ubiquitin-proteasome pathway regulates the levels of p27 protein in a cell. The addition of a proteasome inhibitor is known to prevent the degradation of the p27 protein. Immunoblotting for p27 enabled us to assess the success of the experiment. GAPDH was used as a loading control.

In certain lanes there was a small increase in the levels of the p110δ protein, similarly these lanes also contained a slightly higher level of GAPDH protein suggesting more total protein had been added to these lanes. A steady increase in the levels of p27 protein over 8 h was observed. Therefore we were able to conclude that the ubiquitin-proteasome pathway does not regulate p110δ protein expression.

4.2.5 p110δ protein levels are not altered by stimulation of the B-cell antigen receptor

B cells were purified from the spleen of a wild-type mouse. Anti-IgM antibody (αIgM), at 10 μl/ml final concentration, was added to each plate at certain time points. Two plates remained unstimulated, one of which was lysed at the beginning of the experiment and one at the end to ensure that the 24 h incubation period between the beginning and the end of the experiment did not have an effect the expression of the p110δ protein in the primary cells.

From this set of experiments we can conclude that p110δ expression levels in primary mouse splenic B-cells were not affected by stimulation of the B-cell antigen receptor when treated with anti-IgM antibodies (Figure 4-5).

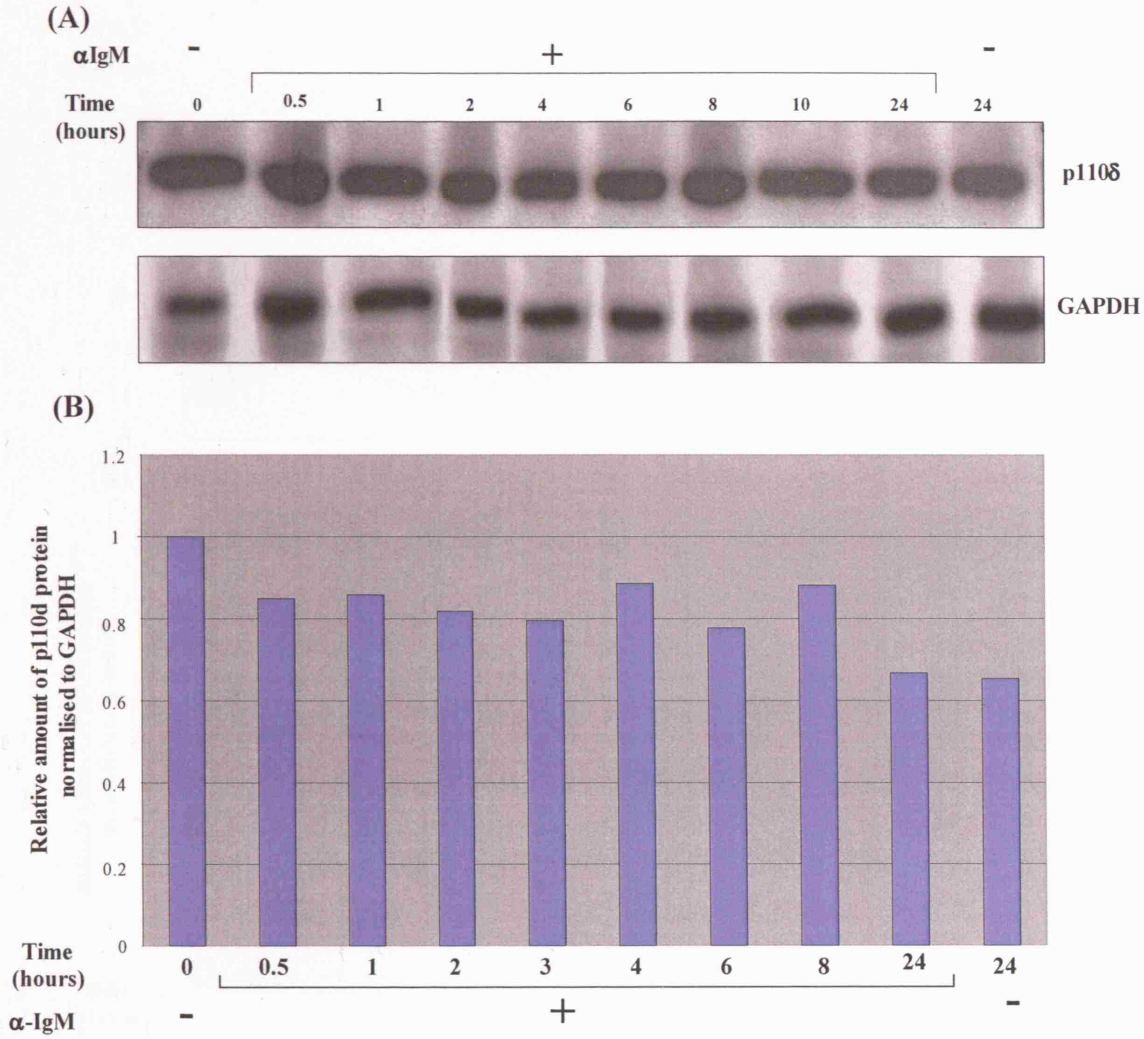

Figure 4-5 The effect of anti IgM antibody stimulation on p110δ expression in splenic B-cells. (A) 100 µg of total cell lysate was resolved on an SDS-PAGE gel and analysed for expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for GAPDH was used to assess protein loading. **(B)** A graph of the p110δ protein signal normalised to the GAPDH protein signal using a densitometer to determine the strength of each of the signals. n = 1

4.2.6 p110δ protein levels are not affected by aldosterone stimulation

Following the protocol used by Tsybouleva *et al.* (199) 10^6 NIH3T3 cells stimulated with 50 µM aldosterone over a period of between 15 min to 4 h and showed no increase in p110δ protein expression (Figure 4-6).

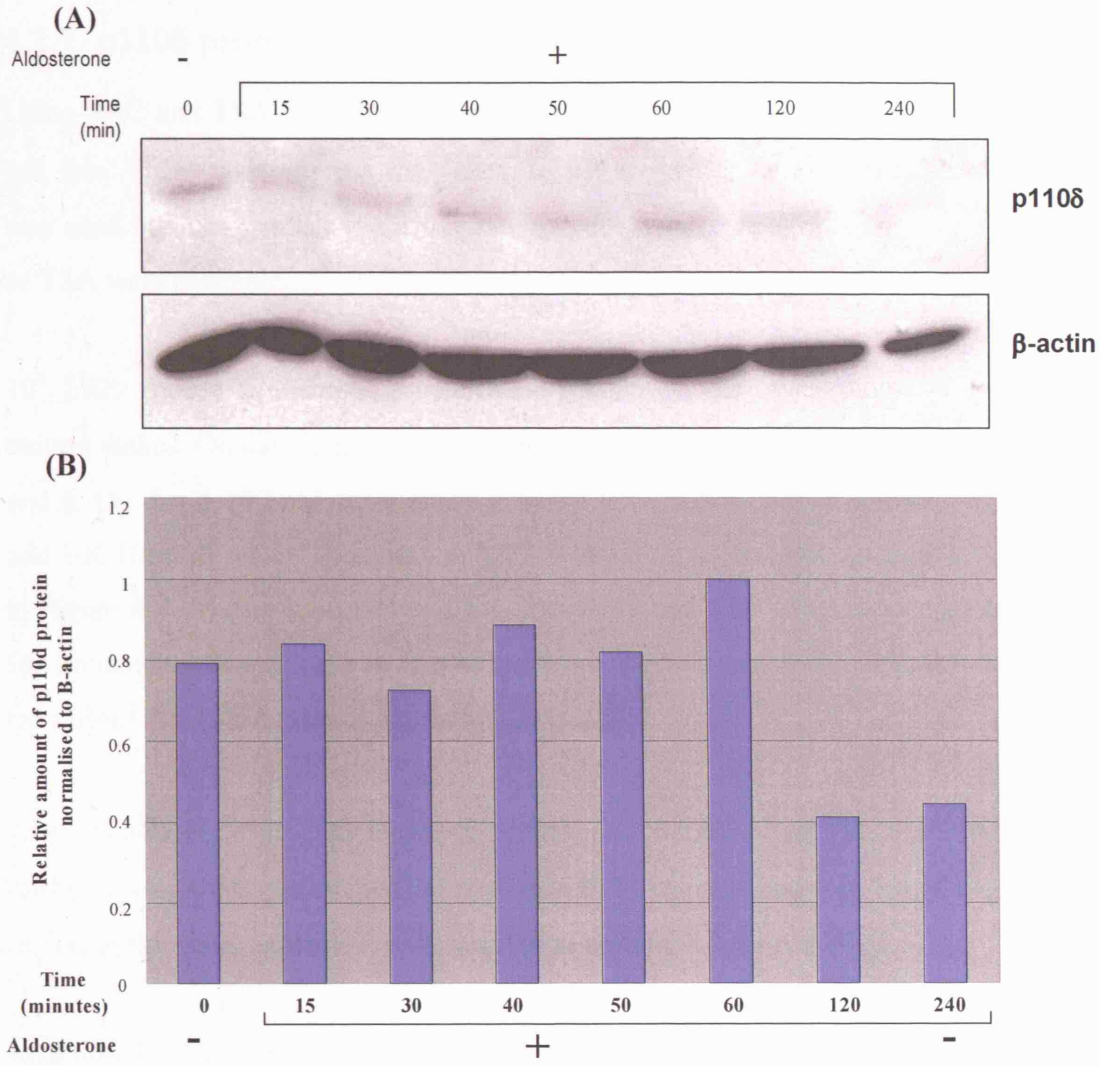

Figure 4-6 The effect of aldosterone stimulation on p110δ expression in NIH3T3 cells. (A) 100 μg of total cell lysate was resolved on an SDS-PAGE gel and analysed for expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for β-actin was used to assess protein loading. **(B)** A graph of the p110δ protein signal normalised to the β-actin protein signal using a densitometer to determine the strength of each of the signals. n = 1

Mineralocorticoids regulate diverse functions important to maintain central nervous system, cardiovascular, metabolic and immune homeostasis. Although they have such a diverse function they only have an affect on mineralocorticoid-responsive cells such as the principal cells of the collecting duct in the kidney. Therefore it would have been desirable to repeat the experiment on cells such as the cardiac fibroblasts, which are known to express the mineralocorticoid receptor (MR).

Taken together, these data indicate that p110δ in NIH3T3 cells is not regulated in an acute manner in response to extracellular stimuli.

4.2.7 p110δ protein levels are not affected by chromatin deregulation

Using 5AC and TSA, we tried to induce p110δ mRNA in the L929 mouse fibroblast cell line. These cells express low levels of p110δ similar to NIH3T3. This cell line was used because published positive controls for induction of specific genes by 5AC or TSA were available.

10^4 L929 mouse fibrosarcoma cells were seeded on day 1 in seven 10 cm tissue culture dishes. On day 2, 5 μM (final concentration) 5AC was added to samples 2,3,5 and 6. On day 5 100 nM (final concentration) TSA was added to samples 4, 5 and 6 and 100 U/ml of recombinant mouse TNF was added to samples 3,6 and 7, as shown in Figure 4-7. The cells were lysed 6 h post TSA and TNF stimulation and prepared for immunoblotting (Figure 4-7) and real-time PCR (Figure 4-8). We assessed both the mRNA levels and protein expression of p110δ.

After treatment with demethylating agents and inhibitors of deacetylation p110δ mRNA levels were determined by real-time PCR using primers in the p110δ coding region of the gene, explained in more detail in section 2.8 (Figure 4-8).

Real time PCR primers for IL6 and IFNβ were used as positive controls. 5AC or TNF did not significantly alter p110δ mRNA levels, whereas TSA induced a 2-fold increase. This induction was not seen at the protein level. Various combinations of these agents did not further increase p110δ mRNA levels. The reason for the discrepancy in p110δ protein expression and mRNA levels is not clear. Experimental error may be a factor as this was a one off experiment. If this is the case the real-time PCR results are more reliable as they were prepared in triplicate.

Taken together, this data suggests that chromatin modulation is not a major level of level of regulation of p110δ gene expression. This experiment also demonstrates that p110δ expression is not upregulated in response TNF-α stimulation which confirms the data discussed earlier in this chapter (4.2.1).

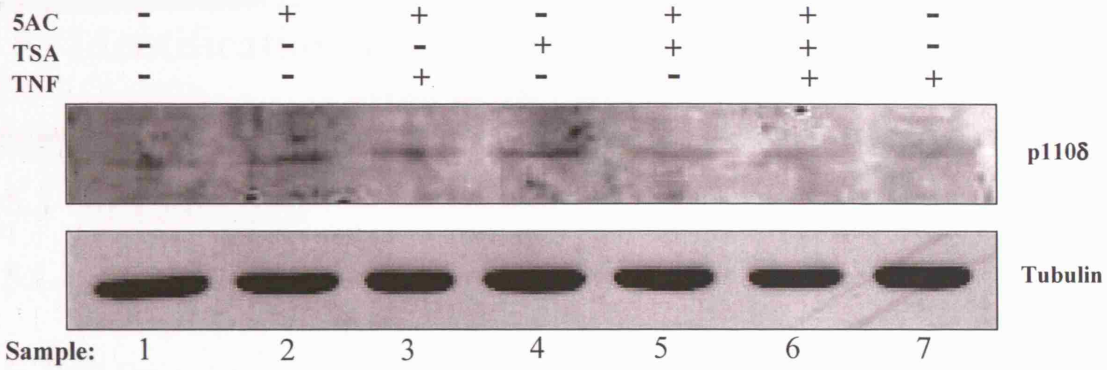

Figure 4-7 The effect of 5AC, TSA and TNF stimulation on p110δ expression in L929 cells. 100 μg of total cell lysate was resolved on an SDS-PAGE gel and analysed for expression of p110δ by immunoblotting using a rabbit polyclonal antibody to p110δ. Blotting for tubulin was used to assess protein loading (experiment performed by Klaartje Kok). n =1

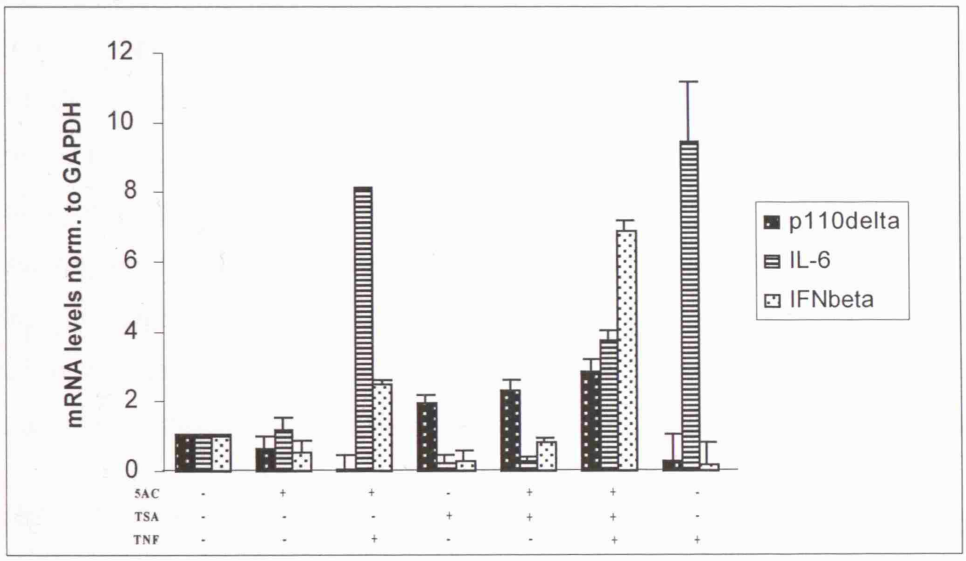

Figure 4-8 p110δ mRNA levels can be induced 2-fold by TSA but are not altered by 5AC treatment. The induction of IL-6 after stimulation with 5AC and TNF and the induction of IFNβ after treatment with TSA, 5AC and TNF confirm that the experiment worked (experiment performed by Klaartje Kok).

5 Identification of three distinct transcriptional start sites in the p110δ gene

5.1 Introduction

5.1.1 ESTs

In 1990 Craig Venter demonstrated a novel method to accelerate gene discovery. At the time, relatively few human genes had been identified and physically mapped to the genome. But molecules called Expressed Sequence Tags (ESTs) offered an efficient way to find genes and explore their functions.

The EST concept arose from the development of complementary DNA (cDNA). When a gene is expressed, its unique sequence of bases is transcribed onto messenger RNA (mRNA). These molecules, which serve as templates for protein synthesis, can be captured from the various tissues in which they are found. Although labile, mRNA molecules can be translated into sturdier complementary DNA (cDNA). By the end of 1990 researchers had created "libraries" of such cDNA molecules. ESTs are cloned segments of cDNA molecules that have been partly sequenced - usually several hundred bases from both ends. These sequenced ends could, in theory, provide information about the location and function of the entire gene they represent (202).

The mRNA from a particular tissue (e.g., liver) is isolated and treated with reverse transcriptase. Reverse transcriptase is a DNA polymerase that uses RNA as its template. Thus it is able to make genetic information flow in the reverse (RNA → DNA) of its normal direction (DNA → RNA). This produces complementary DNA (cDNA). cDNA differs from the normal gene as it lacks the intron sequences. 200–800 nucleotides at both the 5′ and 3′ ends of each cDNA are sequenced and the sequence used to create a database of the organism's genome.

Although controversial when first introduced, ESTs were soon widely used. Not only for rapid identification of new genes, but also for analysing gene expression, gene families, and possible disease-causing mutations.

EST databases were established for various animals, plants, and even fungi. In 1995, Venter and his team published 170,000 ESTs that were used to identify over 87,000 cDNA sequences from various tissues in the human body, over 80 percent of which were previously unknown (203).

The most common method to create the cDNA takes advantage of the run of adenosine nucleotide residues at the 3' end of the mRNA, the poly A tail. Short oligonucleotides containing 12-20 deoxythymidines are used as primers for the reverse transcription of the mRNA to cDNA. The disadvantage of this method is that the reverse transcriptase enzyme may not reach the 5'end of the mRNA thus not identifying the furthest 5'end of the gene when cloned and sequenced. This is most common in large genes. This means that there still exists a bias in these databases due to the original method of construction of the cDNA libraries used to build the databases.

5.1.2 Cloning of p110δ cDNA

Human p110δ was cloned via the isolation of partial PI3K cDNA clones, created via reverse transcriptase-coupled PCR using degenerate primers based on the conserved amino-acid sequence in the PI3K catalytic core domain. This partial p110δ cDNA (~400 nucleotides) was used to screen an oligo-dT primed monocyte cDNA library to isolate full length clones (8). That the cDNA was full length was determined by the presence of the ATG translation start site. This was identified by sequence analysis of potential open reading frames and the position of stop codons 5' of the ATG. However the actual transcription start site of the p110δ mRNA had not been formally identified.

Chantry *et al.* identified the human p110δ gene using a similar method as described above. Resulting PCR products were cloned and sequenced. Data bank searches were performed using the BLAST program, and protein and DNA alignments were made using the Geneworks program. Specific PCR primers were designed based on the sequence of the cloned PCR product and were used to screen the human peripheral blood mononuclear cell (PBMC) library (188).

In an independent search for Ras interacting proteins, using the yeast two-hybrid system Vojtek *et al.* identified two clones related in sequence to p110β, which they called Rip31 and Rip36 (Ras-interacting protein). A PCR-based strategy using gene specific oligonucleotide primers derived from the Rip36 sequence enabled them to isolate full-length cDNA (188). Results from sequence analysis showed that there was 94% homology between the amino acid sequence of the newly isolated cDNA and human p110δ suggesting that they had identified the mouse equivalent of human p110δ.

To determine the 5' transcription start site of the p110δ gene I performed 5'RACE (rapid amplification of cDNA ends) on p110δ mRNA. The 5' untranslated region (5'UTR) of p110δ mRNA that was identified in this manner was subsequently further analysed by RNase protection and standard reverse transcriptase PCR.

5.2 Results

5.2.1 The human and mouse p110δ genes contain three untranslated exons at the 5' end as identified by 5'RACE

Rapid amplification of cDNA ends was performed on three mouse leukocyte cell lines (A20, WEHI 231 and EL4), and three non-leukocyte mouse cell lines (NIH3T3, NmuMG and EMT6). RACE was also performed on untransformed cells including primary MEFs, splenocytes and thymocytes. The human cell lines used for 5' RACE analysis included three non-leukocyte cell lines (MDA-MB-431, HEK 293 and HELA) and two leukocyte cell lines (CCRF-CEM and Jurkat). A description of the cell lines can be found in section 2.2.1. A detailed description of the 5'RACE protocol can be found in section 2.7.

Total RNA was extracted from each of the cell samples using the Ultraspec-II RNA isolation system, described in detail in section 2.4.4.3.

During one step of the RACE protocol, calf intestinal phosphatase (CIP) is added to the samples. CIP removes 5'phosphate group from any degraded mRNA, tRNA, rRNA and DNA, therefore leaving only the full length capped mRNA. This mRNA was treated with tobacco acid pyrophosphatase (TAP). TAP hydrolyzes the bonds in the triphosphate bridge of the cap structure, releasing the cap nucleoside and generating a 5'-phosphorylated terminus on the RNA molecule. The resulting "decapped" 5'-phosphorylated terminus was ligated, using T4 RNA Ligase, to an adapter molecule of known sequence (described in section 2.7.2) with a 3'-hydroxylated terminus.

From a single template molecule, one can produce millions of copies of the PCR product very quickly. The capacity to amplify over a million fold also increases the possibility of amplifying the wrong DNA sequence over one million times. The specificity of PCR is determined by the specificity of the PCR primers. For example, if primers bind to more than one locus, then more than one segment of DNA will be amplified. To control for these possibilities, nested PCR was employed to ensure specificity.

Nested PCR involves two pairs of PCR primers being used on a single locus (Figure 5-1). The first pair amplified the locus as seen in any PCR experiment (Figure 5-1a). The second pair of primers (nested primers) bind within the first PCR product (Figure 5-1b) and produces a second PCR product that will be shorter than the first one. The logic behind this strategy is that if the incorrect DNA was amplified the probability is very low that it would also be amplified a second time by a second pair of primers.

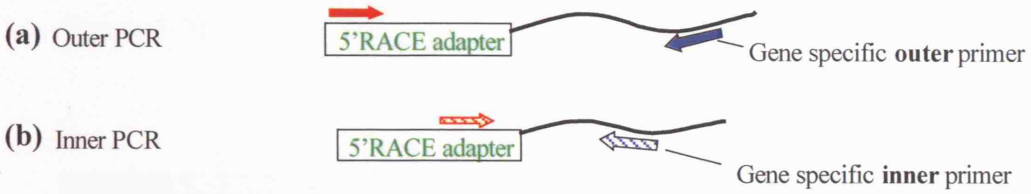

Figure 5-1 Primer positions in a 5'RACE nested PCR assay. The positions of the outer and inner primer-binding sites in the adapter sequence are shown in red. The positions of the outer and inner primer binding sites in the p110δ mRNA are shown in blue.

The forward, outer and inner, PCR primers were complementary to the adapter sequence (shown in red in Figure 5-1). Reverse primers were designed to bind in exon 1 (inner primer) and exon 2 (outer primer) of the human and mouse p110δ mRNA (shown in blue in Figure 5-1 above) (primer sequences can be found in table 2-2). Two consecutive PCRs were performed as described in section 2.3.2 using Taq DNA polymerase. 10 µl of each PCR reaction was run on a 2% high resolution agarose gel containing 1 µg/ml EtBr and visualised on a UV transilluminator .

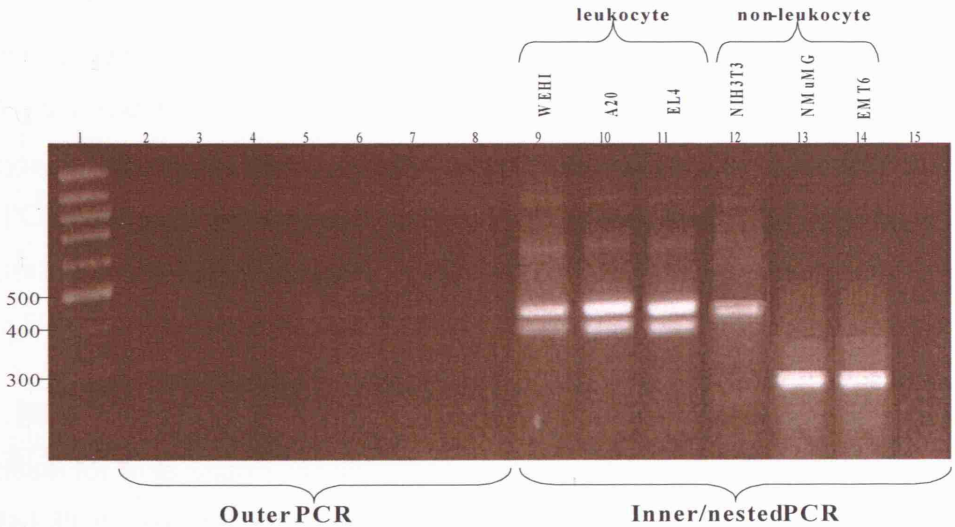

Figure 5-2 Analysis of the 5'RACE products from mouse mRNA by 2% agarose gel electrophoresis and ethidium bromide staining. The first round of PCR, using outer primers, did not show DNA products. Nested PCR analysis using inner primers and 1 µl of the first round reaction resulted in the amplification of two DNA fragments from leukocyte cell lines (lanes 9, 10 and 11) and one DNA fragment from non-leukocyte cell lines (lanes 12, 13 and 14). Lanes 8 and 15 are the negative control samples in which the PCR reaction was carried out without adding the template. n = 3

The three leukocyte cell lines produced two predominant PCR products of about 450 bp and 390 bp. The three non-leukocyte cell lines only produced one PCR product, a 450 bp product in NIH3T3 and a 300 bp product in NMuMG and EMT6 (Figure 5-2). PCR products, similar to those present in the leukocyte cell lines, were identified in mouse primary splenocytes and thymocytes samples. The mouse embryonic fibroblast (MEF) sample showed similarity to the NIH3T3 fibroblast cell line (Figure 5-3).

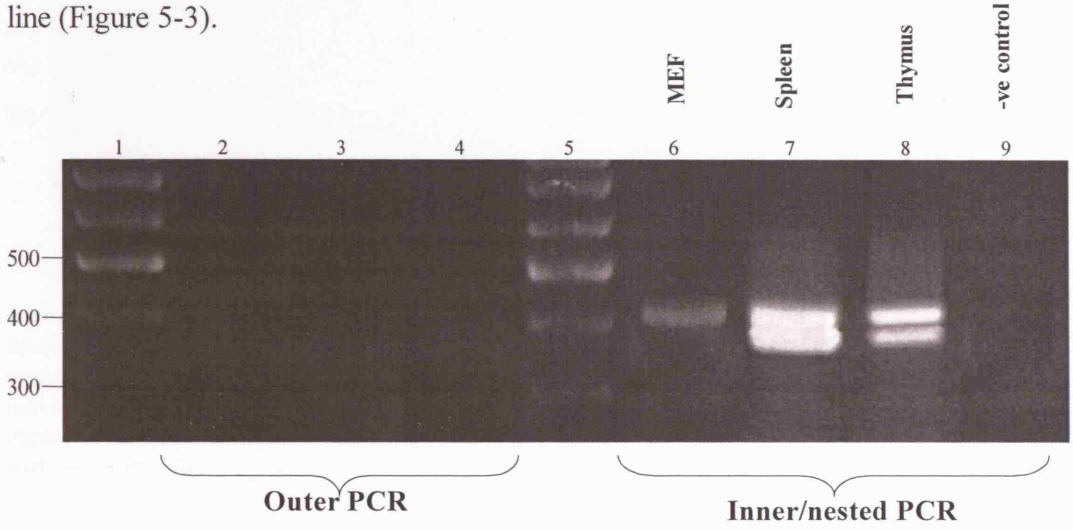

Figure 5-3 Analysis of the 5'RACE products from mouse mRNA by 2% agarose gel electrophoresis and ethidium bromide staining. The first round of PCR did not show any DNA products. Nested PCR analysis using 1 µl of the first round reaction resulted in the amplification one DNA fragment from the MEF sample (lane 6) and two DNA fragments from the splenocytes sample and the thymocytes sample (lanes 7 and 8), respectively. Lane 9 is the negative control sample in which the PCR reaction was carried out without adding the template. n = 2

The same experiment was performed using five different human cell lines. Leukocyte mRNA gave rise to two PCR products of about 350 bp and 520 bp. Instead non-leukocyte mRNA gave rise to a single PCR product. HEK 293 cells gave rise to a single PCR product of the same size as the smaller leukocyte PCR product (~ 350 bp). In contrast, MDA-MB-431 and HELA mRNA gave rise to a smaller 270 bp fragment (Figure 5-4).

Every PCR product was excised from the gel and eluted using the QIAquick purification kit from Qiagen, as described in section 2.3.4. The Taq DNA polymerase amplified PCR products (containing A overhangs) were ligated into the pGEM®-Teasy vector (containing T overhangs as shown in Figure 5-5), using T4 DNA ligase as described in section 2.4.3.

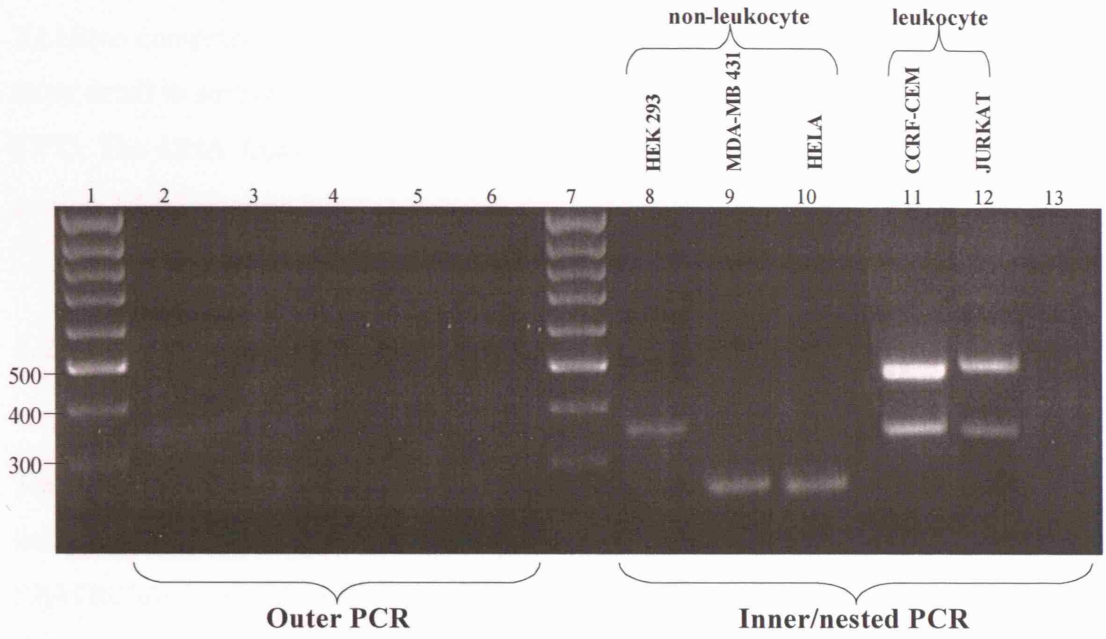

Figure 5-4 Analysis of the 5'RACE products of human mRNA by 2% agarose gel electrophoresis an ethidium bromide staining. The first round of PCR did not show any DNA products. Nested PCR analysis using 1 µl of the first round reaction resulted in the amplification one DNA fragment from the non-leukocyte cell lines (lanes 8, 9 and 10) and two DNA fragments from the leukocyte cell lines (lanes 11 and 12). Lane 13 is the negative control sample in which the PCR reaction was carried out without adding the template. n = 3

The pGEM®-T Easy vector also contains a copy of the *lacZ* gene, which encodes β-galactosidase. Successful cloning of an insert into the vector interrupts the coding sequence of β-galactosidase enabling the recombinant clones to be easily identified by colour screening on agar plates, described in more detail in section 2.4.

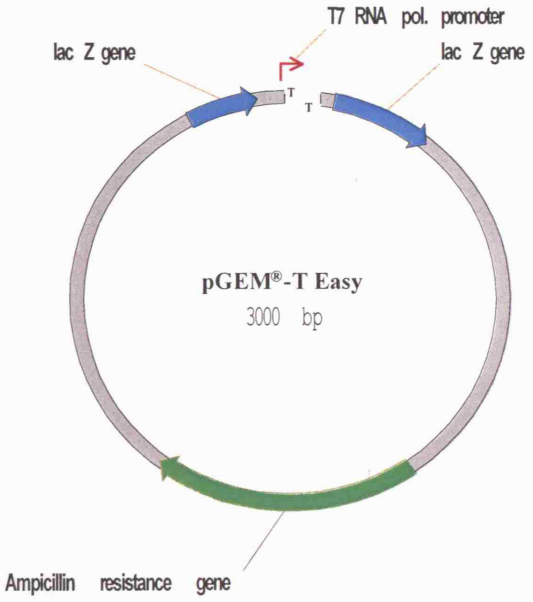

Figure 5-5 pGEM®-T Easy Vector circular map

XL1Blue competent *E.coli* were transformed with the ligation reactions, described in more detail in section 2.4.3(g). The bacteria were grown on Xgal/IPTG agar plates at 37°C. The DNA from the successful transformations was purified, (as described in section 2.4.4.2) before being sequenced.

5.2.2 Sequence analysis of 5'RACE PCR products reveal previously unpublished p110δ sequence 5' of the translation start site.

The DNA extracted from each colony was sequenced, as described in section 2.5, using a primer complementary to the T7 RNA polymerase promoter (**5'** TAATACGACTCACTATAGGG **3'**) (Figure 5-6). The sequences were analysed using the University of California, Santa Cruz, BLAT alignment software - http://genome.ucsc.edu/cgi-bin/hgBlat (described in more detail in section 2.11.2).

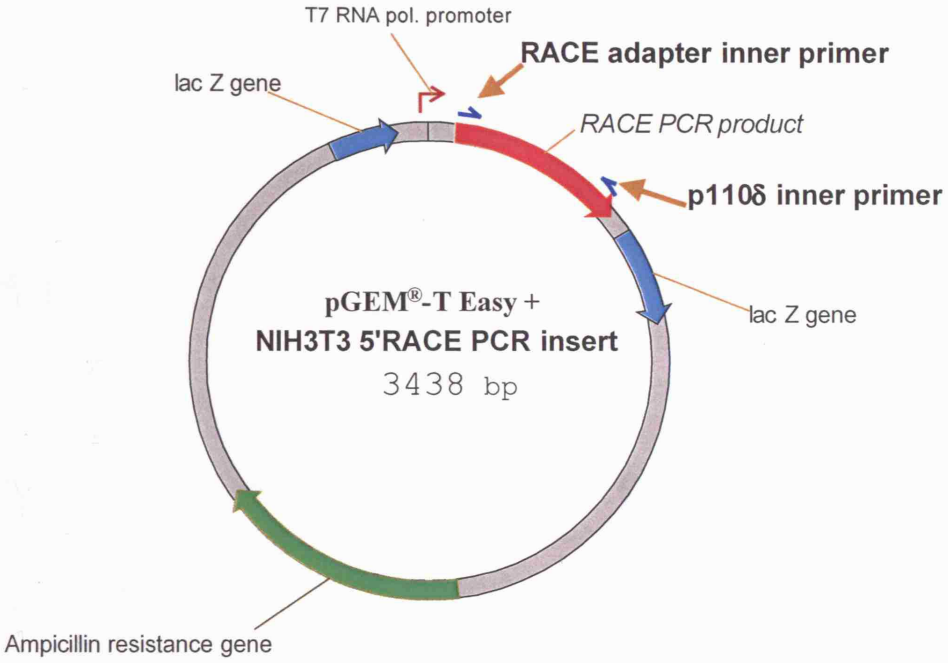

Figure 5-6 pGEM®-T Easy Vector plus the 483 bp PCR product from the 5'RACE of NIH3T3 mouse fibroblast cells. The positions of the forward (RACE) and reverse (p110δ) PCR primers are shown on the vector map. A primer with a sequence complementary to the T7 RNA polymerase promoter (position shown on the map) was used to sequence the RACE products.

Sequence analysis (Figure 5-8) of the 5'RACE PCR products, when aligned to the mouse/human genomic DNA sequence, revealed three distinct mRNA species with different 5' exons (Figure 5-7). These 5' exons are further referred to as exons -2b,

-2a and -1 relative to exon 1 (which contains the ATG translation start codon as defined in refs (8,204)).

All transcripts identified thus far contain exon -1, and two of these further incorporate either exon -2a or exon -2b. A fourth transcript, containing an alternative 5'exon, exon '-2c', was identified in one of the EL4 mouse cDNA clones. It is not clear if this is a true exon, as confirmatory evidence could thus far not be obtained, (explained in sections 5.2.4 and 5.2.5).

A schematic representation of the three distinct transcripts is shown in Figure 5-7A. The genomic location of the p110δ exons is shown schematically in Figure 5-7B.

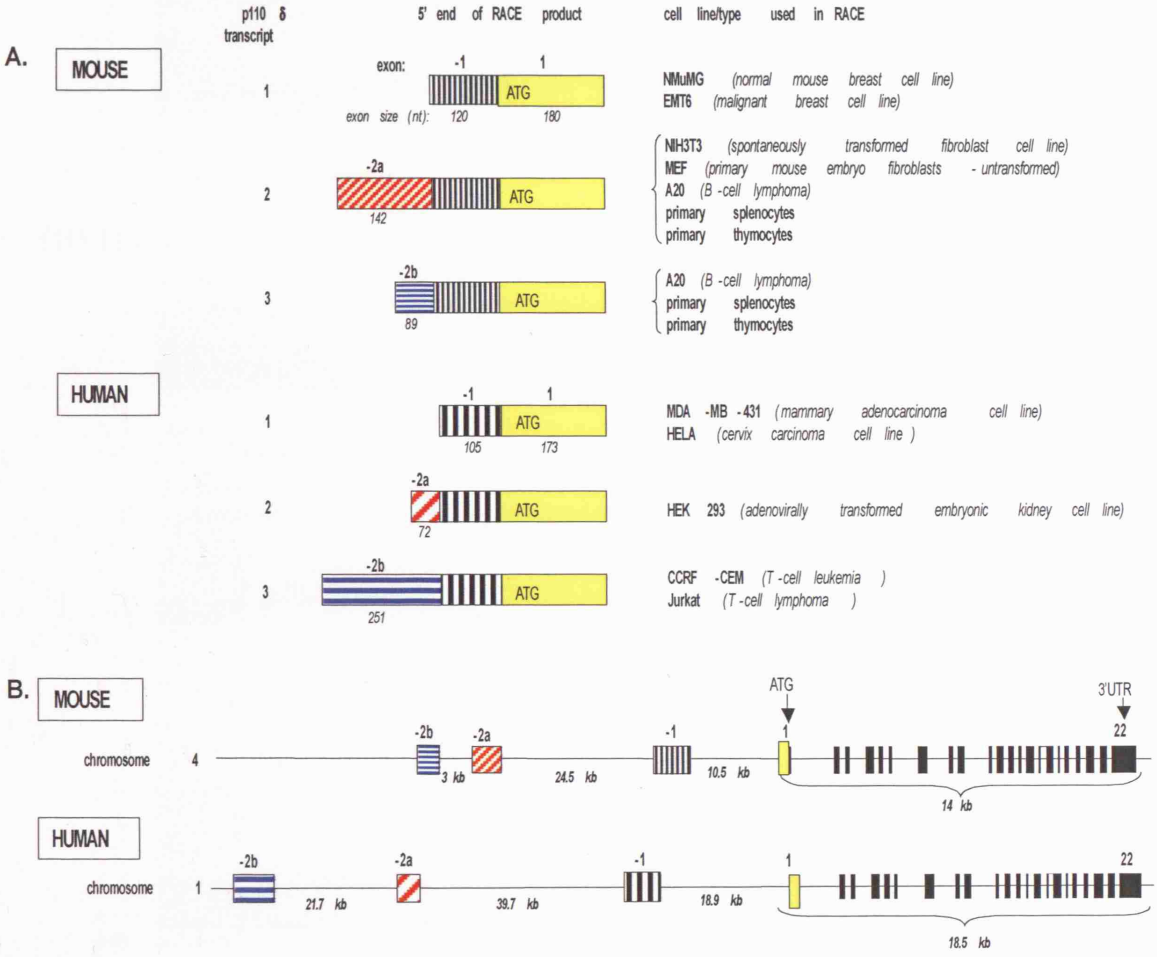

Figure 5-7 A. Diagrammatic representation of the three distinct 5'UTRs of p110δ mRNA transcripts, in mouse and human, as revealed by 5'RACE. **B.** Genomic organisation of the mouse and human p110δ gene.

(A) Mouse –

Transcript 1: exon-1 and exon 1

```
1   TTNGATACGA CTCNCTATAG GGCGAATTGG GCCCGACGTC GCATGCTCCC GGCCGCCATG
61  GCGGCCGCGG GAATTCGATT CGCGGATCCA GACGCTGCGT TTGCTGCTTT GATGAAAACA
121 TCTAAGGAGC TGAGAGCCAG GCAGAAGTGG GATGAAGCCG CTGATGCCAA AGTACCTTTA
181 ATCTCCCAGG CAGAGGGGCC TTGGCTGGTG GTCCTTCTTG GCCCATACCA AACAGGAAAA
241 CAGACCTCAG AGAAGTAACA CAACAGGATG CCCCCTGGGG TGGACTGCCC CATGGAGTTC
301 TGGACCAAAG AGGAGAGCCA GAGCGTGGTT GTTGACTTCT TGCTGCCCAC AGGGGTCTAC
361 TTGAACTTCC CCGTGTTATC ACTATGAATC GCGGCCGCCT GCAGGTCGA
```

Transcript 2: exon -2a, exon -1 and exon 1

```
1   TTNGATACGA CTCNCTATAG GGCGAATTGG GCCCGACGTC GCATGCTCCC GGCCGCCATG
61  GCGGCCGCGG GAATTCGATT CGCGGATCCA GACGCTGCGT TTGCTGCTTT GATGAAACCT
121 GTTATCTGTA GAAAGGAAAC AAAGTGGGAA GTGGAGTGTG CGGACTGTCA GTAGGCGGGC
181 TGTCCCGCTG CGCGCCCCGC CTCTGGCTCA CTCGCGCCTA GCCTTGGGGC TGCCAGCTCC
241 GCCGACCCAG CTGCTGGACA CATCTAAGGA GCTGAGAGCC AGGCAGAAGT GGGATGAAGC
301 CGCTGATGCC AAAGTACCTT TAATCTCCCA GGCAGAGGGG CCTTGGCTGG TGGTCCTTCT
361 TGGCCCATAC CAAACAGGAA AACAGACCTC AGAGAAGTAA CACAACAGGA TGCCCCCTGG
421 GGTGGACTGC CCCATGGAGT TCTGGACCAA AGAGGAGAGC CAGAGCGTGG TTGTTGACTT
481 CTTGCTGCCC ACAGGGGTCT ACTTGAACTT CCCCGTGTTA TCACTATGAA TCGCGGCCGC
541 CTGCAGGTCG A
```

Transcript 3: exon -2b, exon -1 and exon 1

```
1   TTNGATACGA CTCNCTATAG GGCGAATTGG GCCCGACGTC GCATGCTCCC GGCCGCCATG
61  GCGGCCGCGG GAATTCGATT CGCGGATCCA GACGCTGCGT TTGCTGCTTT GATGAAAAGC
121 ATTCCTCCGA GGTCTCTGCA TCAACTCCTG CCCTGTGTGA GTGTCTGTCC TGACTTCCTA
181 AGAAGACGAA CAGTGATGTA GAAGTAACAT CTAAGGAGCT GAGAGCCAGG CAGAAGTGGG
241 ATGAAGCCGC TGATGCCAAA GTACCTTTAA TCTCCCAGGC AGAGGGGCCT TGGCTGGTGG
301 TCCTTCTTGG CCCATACCAA ACAGGAAAAC AGACCTCAGA GAAGTAACAC AACAGGATGC
361 CCCCTGGGGT GGACTGCCCC ATGGAGTTCT GGACCAAAGA GGAGAGCCAG AGCGTGGTTG
421 TTGACTTCTT GCTGCCCACA GGGGTCTACT TGAACTTCCC CGTGTTATCA CTATGAATCG
481 CGGCCGCCTG CAGGTCGA
```

(B) Human

Transcript 1: exon-1 and exon 1

```
1   TTNGATACGA CTCNCTATAG GGCGAATTGG GCCCGACGTC GCATGCTCCC GGCCGCCATG
61  GCGGCCGCGG GAATTCGATT CGCGGATCCA GACGCTGCGT TTGCTGCTTT GATGAAAATA
121 AGGAGTCAGG CCAGGGCGGG ATGACACTCA TTGATTCTAA AGCATCTTTA ATCTGCCAGG
181 CGGAGGGGGC TTTGCTGGTC TTTCTTGGAC TATTCCAGAG AGGACAACTG TCATCTGGGA
241 AGTAACAACG CAGGATGCCC CCTGGGGTGG ACTGCCCCAT GGAATTCTGG ACCAAGGAGG
301 AGAATCAGAG CGTTGTGGTT GACTTCCTGC TGCCCACAGG GGTCTACCTG AACTTCCCTG
361 TGTCCCGTAT CACTATGAAT CGCGGCCGCC TGCAGGTCGA
```

Transcript 2: exon -2a, exon -1 and exon 1

```
1   TTNGATACGA CTCNCTATAG GGCGAATTGG GCCCGACGTC GCATGCTCCC GGCCGCCATG
61  GCGGCCGCGG GAATTCGATT CGCGGATCCA GACGCTGCGT TTGCTGCTTT GATGAAATCG
121 CGCCCAGCGC AGTCGCTCCG AGCGGCCGCG AGCAGAGCCG CCCAGGCCTG CCAGGTGCGC
181 CGGGGGACGA TAAGGAGTCA GGCCAGGGCG GGATGACACT CATTGATTCT AAAGCATCTT
241 TAATCTGCCA GGCGGAGGGG GCTTTGCTGG TCTTTCTTGG ACTATTCCAG AGAGGACAAC
301 TGTCATCTGG GAAGTAACAA CGCAGGATGC CCCCTGGGGT GGACTGCCCC ATGGAATTCT
361 GGACCAAGGA GGAGAATCAG AGCGTTGTGG TTGACTTCCT GCTGCCCACA GGGGTCTACC
421 TGAACTTCCC TGTGTCCCGT ATCACTATGA ATCGCGGCCG CCTGCAGGTC GA
```

Transcript 3: exon -2b, exon -1 and exon 1

```
1   TTNGATACGA CTCNCTATAG GGCGAATTGG GCCCGACGTC GCATGCTCCC GGCCGCCATG
61  GCGGCCGCGG GAATTCGATT CGCGGATCCA GACGCTGCGT TTGCTGCTTT GATGAAAGAG
121 GCGCCCAGGA CACCACGAGG TTGGGAGAGG AGTGTGCTTT GCACTCTGCA CTCTCCAGCT
181 CAGAGTAGCT GAGGATGCTA GGGAGACTGG GGGCTGCCAC GCGGGGGGAG CAGAGGAGTC
241 CTCAGTGCGG TCCTCACTCT GAAAACAGCA TTTTTCCCTA GGCTTGAGAA TGCCTCCGGG
301 ATTGGACTGA TCCAGGCTGG ACCCACGTCT GTCTGGTGAT ACCAGGGGCA GAGGGACCAC
361 TCTGACAGAT AAGGAGTCAG GCCAGGGCGG GATGACACTC ATTGATTCTA AAGCATCTTT
421 AATCTGCCAG GCGGAGGGGG CTTTGCTGGT CTTTCTTGGA CTATTCCAGA GAGGACAACT
481 GTCATCTGGG AAGTAACAAC GCAGGATGCC CCCTGGGGTG GACTGCCCCA TGGAATTCTG
541 GACCAAGGAG GAGAATCAGA GCGTTGTGGT TGACTTCCTG CTGCCCACAG GGGTCTACCT
601 GAACTTCCCT GTGTCCCGTA TCACTATGAA TCGCGGCCGC CTGCAGGTCG A
```

Figure 5-8 Sequence of the three mRNA transcripts from mouse and human 5'RACE cloned into the pGEM® T Easy vector. (A) Sequence of mouse 5'RACE products **(B)** Sequence of human 5'RACE products. Grey sequence – pGEM® TEasy cloning vector. Green sequence – RACE adapter. Blue sequence – exon -2b. Red sequence – exon -2a. Black sequence – exon -1. yellow sequence – exon 1. Yellow boxed sequence – p110δ internal reverse PCR primer. ATG translation start site (as predicted by (204)) is highlighted in pink.

An in-frame stop codon (Figure 5-9: TAA – turquoise) present immediately upstream of the ATG start codon (Figure 5-9– pink) in both mouse and human p110δ mRNA ruled out the possibility that the newly identified upstream exons are translated as part of the p110δ protein.

Figure 5-9 An in-frame stop codon, in mouse and human p110δ mRNA, immediately upstream of the ATG translation start site ruled out the possibility that the 5'exons are translated as part of the p110δ protein. Black sequence – exon -1; Yellow sequence – exon 1; ATG translation start site is highlighted in pink and the stop codon is highlighted in turquoise.

There is 61.5% sequence homology between exons 1-22 in the mouse and human p110δ gene, as determined by BLAST 2 Sequences (NCBI). Exon 22 of the mouse and human p110δ gene is 1608 bp and 2015 bp, respectively. Only 271 bp and 286 bp of this exon in mouse and human, respectively, are translated. The remainder makes up the 3' UTR of the p110δ gene. The 14.5% homology between exon 22 of the mouse and human p110δ gene is within the coding region of this exon (Table 5-1).

Exons of the p110δ gene	% Homology between human and mouse
-2b	none
-2a	none
-1	60.8
1	92
2-21 (coding region)	89
22	14.5

Table 5-1 Analysis of the homology between the mouse and human p110δ exons using the BLAST 2 sequence alignment tool. No homology was identified between the mouse and human exons -2a and -2b. High sequence homology was recorded between exon -1 of the mouse and human p110δ gene. The coding exons 1-22 are highly homologous. 80% of exon 22 is non-coding reducing the overall percent homology of this exon between the mouse and human.

Standard alignment of individual exons in the 5' UTR of the human and mouse p110δ gene revealed 60.8% homology between the mouse and human exon -1 but no homology between the other 2 upstream exons, exons -2a and -2b (Table 5-1). From this analysis it is evident that non-coding exons are not as highly conserved as coding exons.

No 5'UTR sequences were present in the mouse p110δ cDNA sequence originally submitted to GenBank (accession number U86587; Refs.(188)). The human p110δ cDNA sequence that Bart Vanhaesebroeck submitted to Genbank (accession number Y10055; clone o9 described in Ref. (8)) contains 196 nucleotides 5' of the ATG start codon (Figure 5-10). Based on the 5'RACE analysis described here, it is now clear that this clone represents transcript 2, containing exon -2a (missing 21 bases at the 5' end), exon -1 and exon 1.

HUMAN

Transcript 2: exon -2a, exon -1 and exon 1

```
1   TCGCGCCCAG CGCAGTCGCT CCGAGCGGCC GCGAGCAGAG CCGCCCAGCC CTGCCAGCTG
61  CGCCGGGGGA CGATAAGGAG TCAGGCCAGG GCGGGATGAC ACTCATTGAT TCTAAAGCAT
121 CTTTAATCTG CCAGGCGGAG GGGGCTTTGC TGGTCTTTCT TGGACTATTC CAGAGAGGAC
181 AACTGTCATC TGGGAAGTAA CAACGCAGGA TGCCCCCTGG GGTGGACTGC CCCATGGAAT
241 TCTGGACCAA GGAGGAGAAT CAGAGCGTTG TGGTTGACTT CCTGCTGCCC ACAGGGGTCT
301 ACCTGAACTT CCCTGTGTCC CG
```

Figure 5-10 cDNA sequence of human p110δ transcript 2. Red sequence – exon -2a. Black sequence – exon -1. Yellow sequence – exon 1. The sequence of the o9 clone of p110δ, upstream of the ATG, originally submitted to Genbank (Y10055; ref. (8)) is highlighted in BOLD.

During cloning of human p110δ, a cDNA (called o5) was isolated that contains exon -2a and 1, but lacked exon -1 (Bart Vanhaesebroeck, unpublished results), indicating that some p110δ transcripts may lack exon -1. One such transcript was also found in the Riken database (4M- Figure 5-12).

Specific nucleotide sequences at the 5' end (splice donor) and the 3' end (splice acceptor) of each intron are recognised by splicing machinery to ensure the accurate removal of the introns.

Thus far, no RACE products that incorporate both exon -2b and -2a have been found. This indicates that transcript 3 (incorporating exons -2b/-1) arises from -2b/-2a/-1 transcripts by splicing out of exon -2a. That this is possible is indicated by the

presence of a splice donor at the 5' end of the intron between exon -2b and exon -2a and an acceptor at the 3' end of the intron between exon -2a and exon -1 in both the mouse and human p110δ gene (Figure 5-11 and Table 5-2). The absence of an acceptor at the 3' end of the intron between exon -2b and exon -2a (Table 5-2 column 8) makes it highly unlikely that exon -2b and exon -2a are spliced to each other. This helps to confirm the results obtained from the 5'RACE experiment that exon -2b is spliced onto exon -1, therefore removing exon -2a from transcript 3.

The first transcribed exon of mRNA is not known to be spliced out in any gene. For example, transcript 1, containing exon -1 alone, could not be created from either transcript 2 or transcript 3 having spiced out exon -2a or exon -2b, respectively. Taken together, these observations suggest the presence of 3 distinct promoter regions, upstream of exons -2b, -2a and -1, respectively.

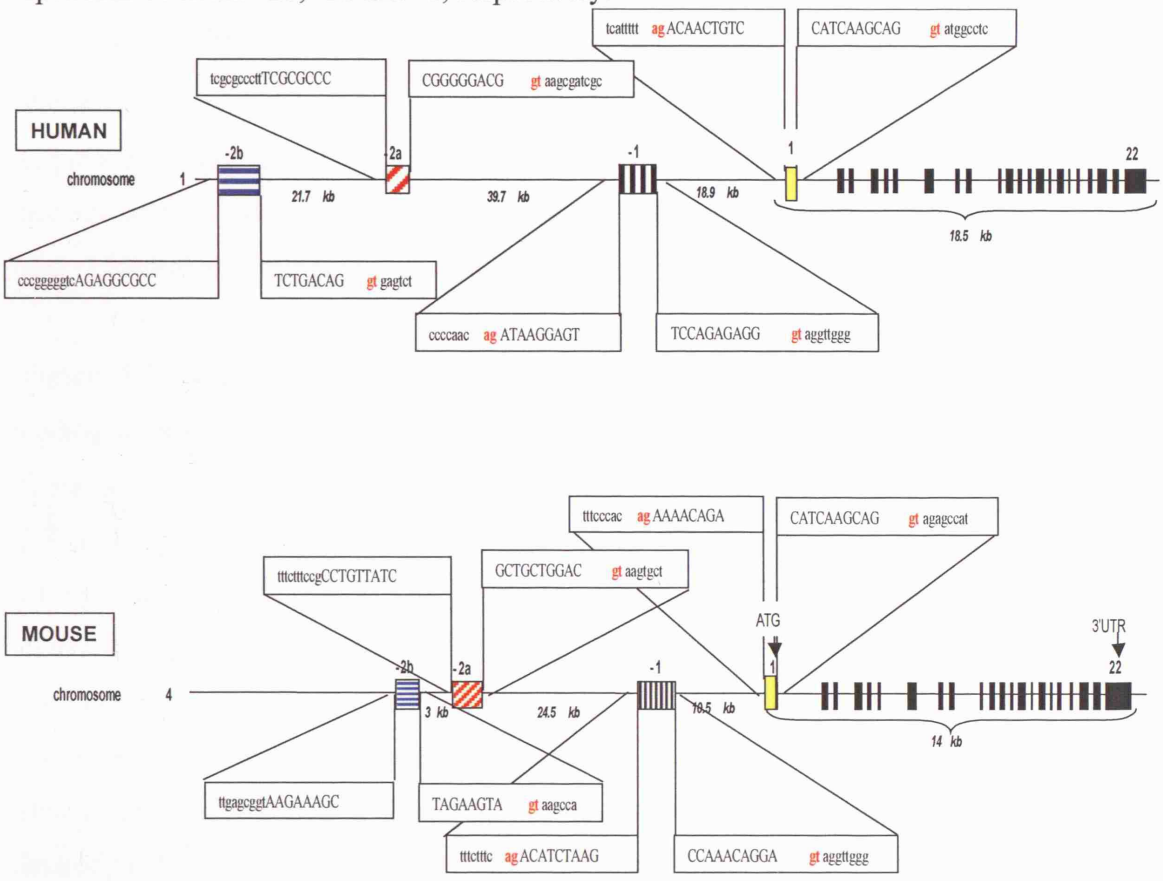

Figure 5-11 A diagrammatic representation of the genomic organisation of the mouse and human p110δ gene.

Species	Exon		Exon Size	5' splice donor		Intron size	3'splice acceptor	
				exon sequence	5' intron sequence		3'intron sequence	exon sequence
								-2a
Human	-2b	5'UTR	211bp	...ACTCTGACAG	gtgagtctag	21.5kb	tcgcgcccttt	TCGCGCC...
Mouse	-2b	5'UTR	80bp	...ATGTAGAAGT	gtaagccaaa	27.4kb	tttctttccg	CCTGTTATC...
								-1
Human	-2a	5'UTR	73bp	...GCCGGGACG	gtaagcgatcgc	40kb	tccccaacag	ATAAGGAGT...
Mouse	-2a	5'UTR	122bp	...TGCTGGACCG	gtaagtgct	24.35kb	tttctttcag	ACATCTAAG...
								1
Human	-1	5'UTR	93bp	...TTCCAGAGAG	gtaggttggg	18.8kb	tcatttttag	GACAACTGT...
Mouse	-1	5'UTR	100bp	...TACCAAACAG	gtaggttggg	10.8kb	tttcccacag	GAAAACAGA...
								2
Human	1	ATG exon	171bp	...CATCAAGCAG	gtatggcctc	4.95kb	gtccctccag	CTGCTGTGG...
Mouse	1	ATG exon	173bp	...CATCAAGCAG	gtagagccat	2.91kb	tctctcccag	GTGCTGTGG...

Table 5-2 Splice donor and acceptor sites in between the mouse and human p110δ 5'untranslated exons. The 5'splice donor (gt) and 3'splice acceptor (ag) sites present in the intron sequences between the 5'untranslated exons of p110δ mRNA.

5.2.3 Evidence that the three 5'untranslated exons of p110δ, identified by RACE, are not cloning artifacts

While our work was in progress, genome database information derived from EST sequencing and genome-wide determination of the full-length mRNAs had accumulated, providing independent confirmation of the presence of exons -2b, -2a and -1 identified in our study.

Figure 5-12 summarizes the currently available information on potential 5' non-coding exons of the p110δ gene. This database information contains the cell line/ tissue sample from which the EST was identified. In 2002 there was a large influx of ESTs published by the RIKEN consortium, upon the publication in Nature "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs" (205). The sources of the mRNA samples were leukocyte-rich organs such as the spleen, the thymus and the bone marrow. The primary cells that were analysed were largely dendritic leukocyte cells. Details of the mRNA sources of individual transcripts can be found in Table 5-3. This therefore created a bias toward mouse leukocyte EST data, which most likely explains the high frequency of sequences containing exon -2b which, as will be explained in section 5.2.5, is only found in leukocytes (Figure 5-12).

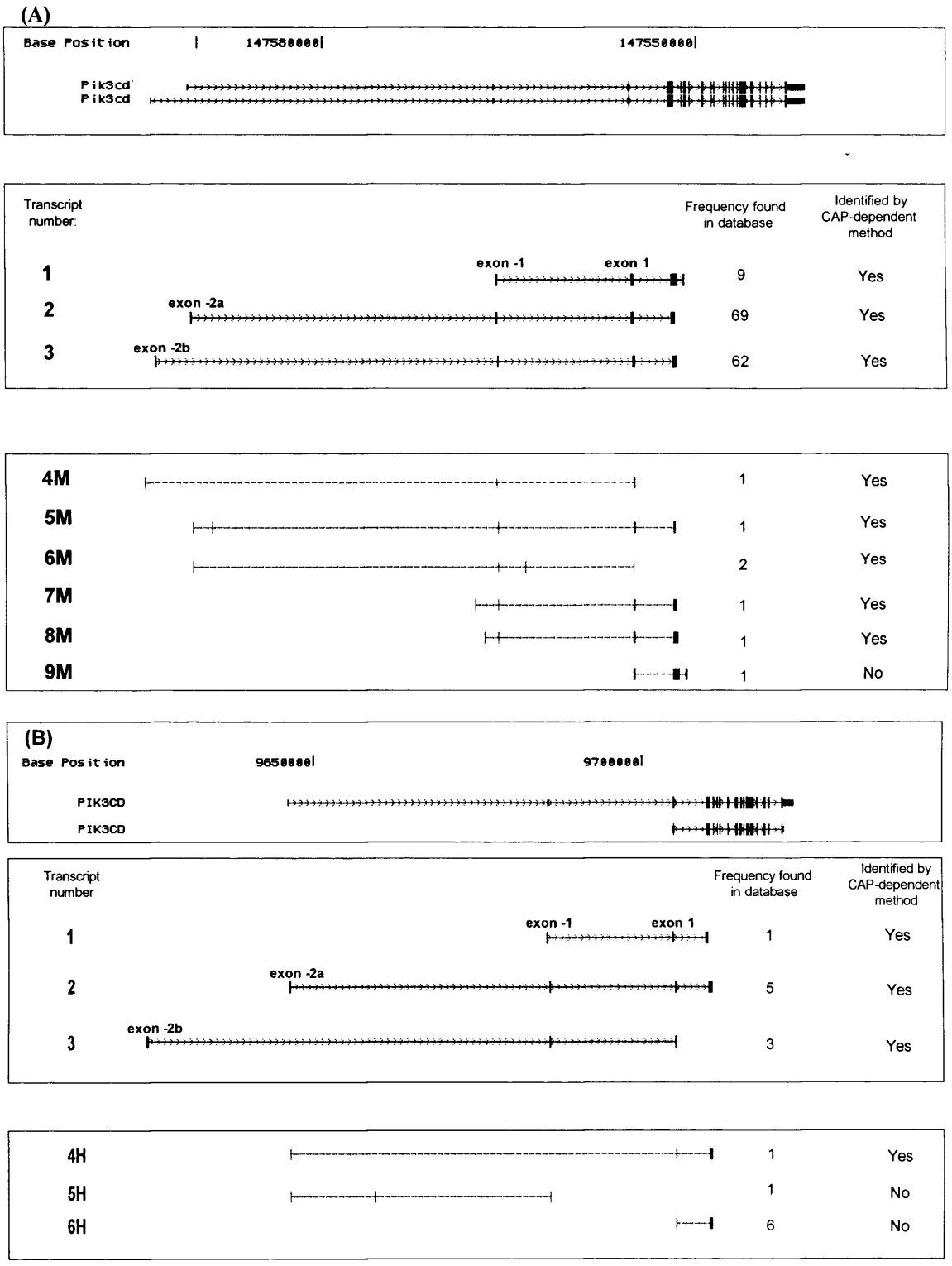

Figure 5-12 p110δ exons as identified by the UCSC Genome Bioinformatics Group database (http://www.genome.ucsc.edu May 2005) on the published mouse and human p110δ gene and p110δ ESTs containing the 5'UTR. The figure summarises all of the published mouse **(A)** and human **(B)** p110δ ESTs in the UCSC Genome database containing part or all of the 5'UTR. The frequency of the different transcripts in the database and whether they were identified using a CAP-dependent or CAP-independent method is also indicated. Filled boxes represent exons and the small arrows along the lines between the exons represent the introns and the direction in which the gene is read. *Top box* in (A) and (B) are known genes based on Swiss Prot, Trembl, mRNA and RefSeq data. *Middle box* in (A) and (B) are transcripts also identified using the Ambion 5'RACE kit.

Evidence for transcripts in addition to transcripts 1, 2 and 3 identified in our study, is present in databases. However, the configuration of these putative transcripts is very different between human and mouse p110δ. We have named these transcripts either M (mouse) or H (human) with a numbering of additional transcripts that does not refer to corresponding transcripts in human and mouse (for example transcript 4H is not comparable to transcript 4M). Moreover, we have thus far not identified any of these transcripts by 5' RACE (50 independent sequences tested in total; detailed data not shown). It therefore remains to be seen whether these additional transcripts are real, or are cloning or sequencing artefacts.

(A) MOUSE

Transcripts	Source of transcript (organ/tissue or cell line)	Institute	CAP dependent isolation method	frequency of transcript
1	Activated spleen	Riken	yes	5
1	Adult inner ear	Riken	yes	3
1	Whole brain - embryos 13.5-17.5 dpc	NIH-MGC	no	1
2	Activated spleen	Riken	yes	23
2	NOD derived CD11c +ve dendritic cells	Riken	yes	18
2	B6 derived CD11 +ve dendritic cells	Riken	yes	14
2	Enriched bone marrow macrophage	Riken	yes	9
2	Adult male bone	Riken	yes	1
2	Melanocyte	Riken	yes	1
2	Whole body - 17.5 day embryo	Riken	yes	1
2	Osteoclast like cell	Riken	yes	1
2	Salivary gland	NIH-MGC	no	1
3	B6 derived CD11 +ve dendritic cells	Riken	yes	28
3	NOD derived CD11c +ve dendritic cells	Riken	yes	17
3	Enriched bone marrow macrophage	Riken	yes	6
3	Mouse 8cell embryo cDNA library	NIA	no	6
3	Adult male aorta and vein	Riken	yes	1
3	Embryonic germ cell cDNA library	NIA	no	1
3	Adult inner ear	Riken	yes	1
3	Adult male thymus	Riken	yes	1
3	Neonate thymus (3 days old)	Riken	yes	1
4M	NOD derived CD11c +ve dendritic cells	Riken	yes	1
5M	Enriched bone marrow macrophages	Riken	yes	2
6M	Whole brain - embryo 15.5 dpc	NIH-MGC	no	1
6M	Whole brain-new born 1,5 and 15 days	NIH-MGC	no	1
7M	Activated spleen	Riken	yes	1
8M	NOD derived CD11c +ve dendritic cells	Riken	yes	1
9M	Whole eye - embryos 12.5-14.5 dpc	NIH-MGC	no	1
			Total	**160**

(B) HUMAN

Transcripts	Source of transcript (organ/tissue or cell line)	Institute	CAP dependent isolation method	frequency of transcript
1	Germinal centre B cell	NCI	no	1
1	Pooled germ cell tumours	NCI	no	1
1	Adult	RZPD	no	1
2	Adult	RZPD	no	3
2	Pediatric B cell acute lymphoblastic leukaemia	TCC - Riken	yes	1
2	Embryonic stem cells	RMGC	yes	1
3	Pediatric acute myelogenous leukaemia	TCC - Riken	yes	1
4H	Pediatric acute myelogenous leukaemia	TCC - Riken	yes	1
5H	Pediatric B cell acute lymphoblastic leukaemia	TCC - Riken	yes	1
6H	Germinal centre B cell	NCI	no	2
6H	Pooled foetal lung, testis and B cell	NCI	no	1
6H	Colon	NCI	no	1
6H	Adult genomic DNA	Sanger	no	1
6H	Marrow	LICR - brazil	no	1
			Total	17

Table 5-3 The source of p110δ transcripts (organ, tissue or cell line) in the UCSC Genome Bioinformatics database and the frequency at which they appear. Column 4 shows whether the method used to isolate the mRNA was CAP-dependent, which provides some indication for the likelihood of full length mRNAs (CAP-dependent methods are more likely to identify the 5' end of the transcript).

5.2.4 Verification of the presence of transcripts containing exon-2a or exon-2b by RNase protection

The RNase protection assay is a sensitive technique with which the presence and quantity of specific RNA in total cellular RNA or polyA-selected mRNA can be assessed. Complementary, radioactively labeled *in vitro* transcript probes are incubated with cellular RNA of choice, allowing the probe and the target RNA to hybridize in solution. The mixture is then treated with RNase to degrade the remaining single stranded RNA. The hybridized double stranded portion of the probe is protected from digestion and can be visualized by electrophoresis on a denaturing polyacrylamide gel followed by autoradiography (described in section 2.9). Since the protected fragments are analysed by high-resolution polyacrylamide gel electrophoresis, the RNase protection assay can be used to accurately map mRNA features such as initiation and termination. If the probe is in molar excess with respect to the target RNA, then the signal is proportional to the amount of complementary RNA in the sample.

pGEM® T Easy, the vector in which the 5'RACE products were cloned, contains both an SP6 polymerase promoter and a T7 polymerase promoter, one either side of

the cloning site. One clone containing each of the newly identified exons was used as a template for probe preparation (Figure 5-13).

The orientation of the insert in the vector is of great importance. The RNA probe must be transcribed so that it is complementary to the RNA of interest. The vector must be linearised at the opposite end of the insert to the promoter to prevent transcription of the pGEM® T Easy vector sequence following the p110δ insert sequence. Therefore the choice of RNA polymerase and restriction enzyme to linearise the template plasmid are vital.

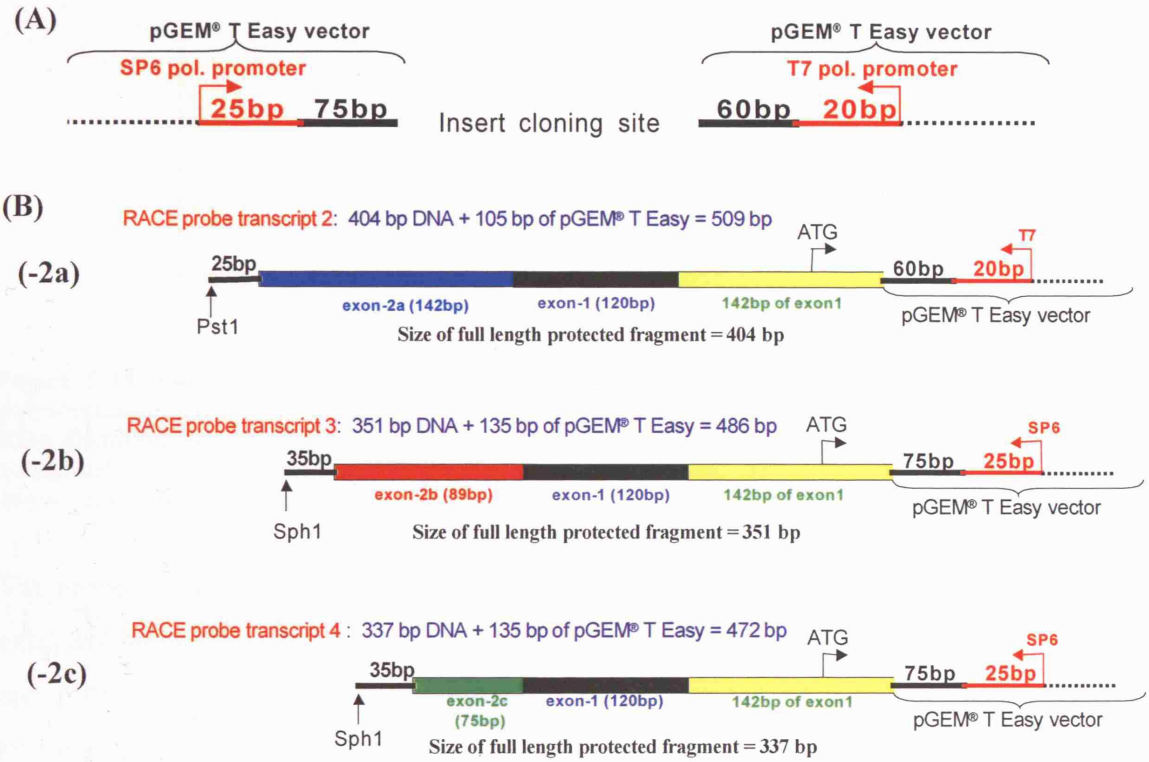

Figure 5-13 Schematic representation of three of the mouse p110δ 5'RACE transcripts subcloned in the pGEM® T Easy vector. (A) The distance between the two promoter sites in the pGEM® T Easy vector and the cloning site. **(B)** The orientation in which the 5'RACE PCR products were cloned into the pGEM® T Easy vector. The expected sizes of the -2a, -2b and -2c probes, if fully protected, are 404 bp, 315 bp and 337 bp, as indicated under each of the respective diagrams.

The vectors were linearised and the RNA reverse transcribed using the appropriate restriction enzymes and RNA polymerase, described in detail in section 2.9.2.1 (Figure 5-13). The now radioactively labelled probes were run on a denaturing polyacrylamide gel to analyse the outcome of the transcription reaction (Figure 5-14).

All of the probes were successfully transcribed. All sizes were in-line with the expected fragment sizes of the radioactively labelled probes: -2a: 509 nucleotides; -2b: 486 nucleotides, -2c: 472 nucleotides and the β-actin: 334 nucleotides (Figure 5-14).

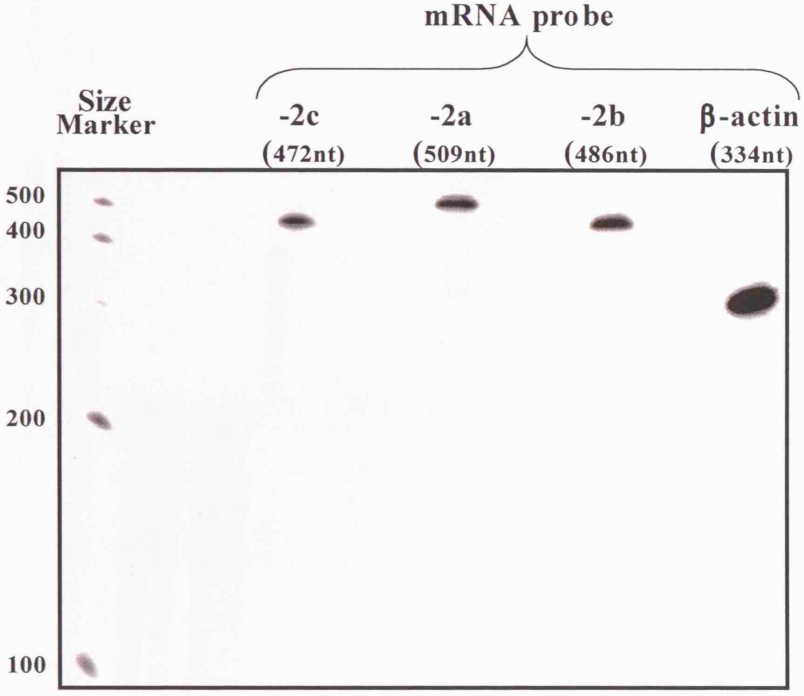

Figure 5-14 The reverse transcribed radioactively labelled probes run on a denaturing polyacrylamide gel next to an RNA size marker. The size marker contained radioactively labelled RNA fragments of 100, 200, 300, 400 and 500 nucleotides. All the probes were successfully transcribed. All sizes were in-line with the expected fragment sizes: -2c – 472 nt, -2a – 509 nt, -2b – 486 nt, and the control probe β-actin – 334 nt. n = 5

The probes were eluted from the gel, purified and incubated for 16 h with total RNA extracted from one non-leukocyte (NIH3T3) and three leukocyte (WEHI 231, EL4 and A20) mouse cell lines. The unprotected single-stranded RNA was then digested by RNase, described in section 2.9.2.3. Total RNA from yeast was used as a positive control for the function of the RNase. Yeast do not have class I PI3Ks therefore any non-specific protection of the p110δ mRNA would be evident in the yeast control lane. β-actin was used as a positive control to assess the quality of the sample RNA. A specific β-actin signal is expected if the cell line mRNA is of good quality.

The double-stranded protected RNA was precipitated and resuspended in gel loading buffer before being denatured and run on a denaturing polyacrylamide gel. The gel was transferred to filter paper and exposed to X-ray film (Figure 5-15).

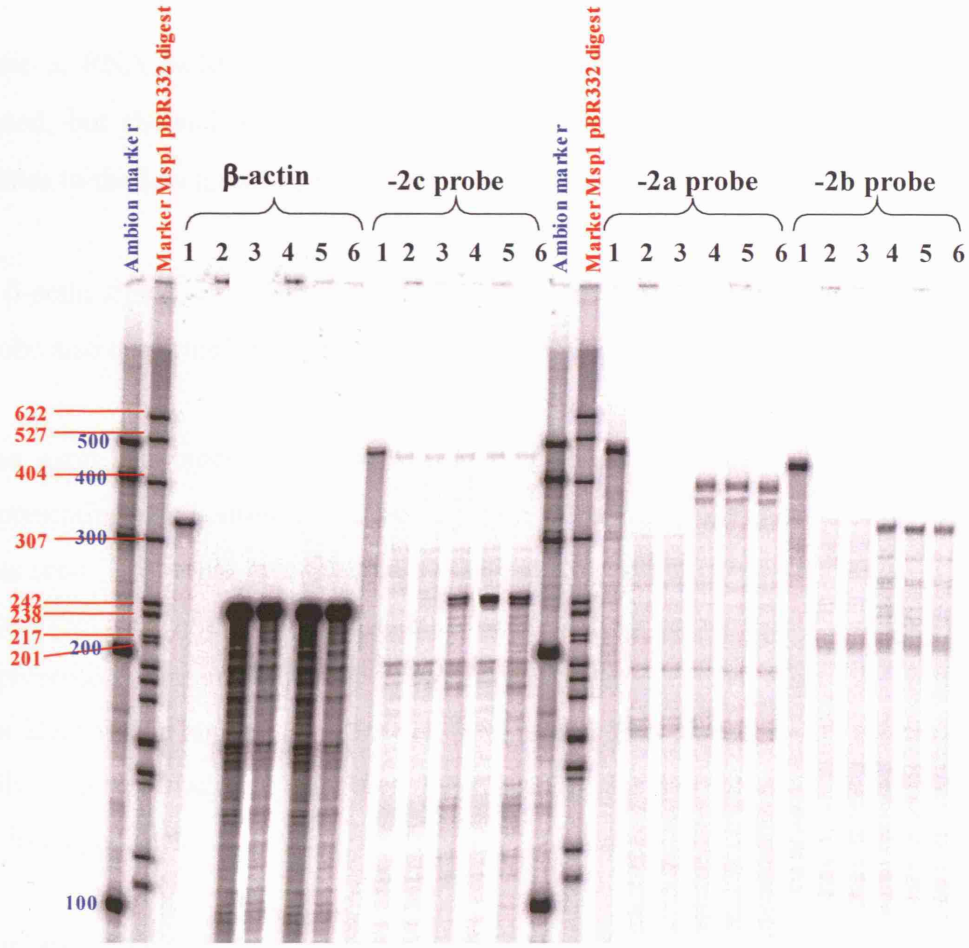

Figure 5-15 Detection of upstream exons in mouse cell lines by RNase protection. 5 different RNA samples were incubated with 4 different RNA probes (β-actin, -2c, -2a and -2b, schematically shown in Figure 5-13. For sequences see Figure 5-8). Samples in numerical order: 1) RNA probe alone 2) Yeast RNA 3) NIH3T3 RNA 4) WEHI 231 RNA 5) EL4 RNA 6) A20 RNA. n = 5

Lane 1, for each probe, was unhybridised, undigested sample used to assess the integrity of the radioactively labeled probes. The signals are in-line with the expected size of the probes: β-actin: 334 nt, -2c: 472 nt, -2a: 509 nt and -2b: 486 nt.

Lane 2, each probe was hybridized with yeast RNA. Yeast does not express Class IA PI3K therefore yeast RNA was used as a negative control to assess any non-specific binding properties of the probes. The absence of a signal in lane 2 confirmed that none of the probes bound to non-specific sequences in yeast RNA. Complete digest of the probes in this lane also recognized the activity of the RNase enzyme.

Lane 3, RNA isolated from NIH3T3 cells did not protect any of the p110δ probes tested, but showed a good signal with the β-actin control probe. This most likely relates to the low level of p110δ mRNA expression in this fibroblast cell line.

A β-actin signal, of around 240 nt, present in the samples hybridized with the β-actin probe also confirmed the integrity of the WEHI 231, EL4 and A20 RNA samples.

The exon -2c probe was designed to protect 337 bp of p110δ mRNA, a signal representing this configuration was not seen (Figure 5-15). Instead a signal of 260 bp was seen. This could be explained by the transcription of exon 1 and exon -1, without exon -2c (see Figure 5-13(B)). It is important to recall that exon -2c was only represented once out of the 25 5'RACE products sequenced from EL4 cells. It was not identified in any of the other cell lines tested. This together with the absence of a fully protected fragment using the -2c probe suggests that the exon -2c may have been a cloning artifact.

The exon -2a and -2b probes were designed to protect a 404 nucleotide and 351 nucleotide length of p110δ mRNA, respectively. The signals seen in lanes containing the leukocyte cell lines (4, 5 and 6) correspond to the expected sizes, thus demonstrating that transcripts containing exon -2a and exon -2b are expressed in these cell lines.

Transcript 3, containing exon -2b, -1 and 1, and transcript 2, containing exon -2a, -1 and 1 were identified by 5'RACE in leukocyte cell lines and in mouse organs abundant in leukocytes. However, non-leukocyte cell lines appeared only to express transcript 2. RNase protection confirmed that the three leukocyte cell lines tested expressed transcript 2 and 3.

However the assay was not sensitive enough to detect a signal from the non-leukocyte cell line tested.

5.2.5 The p110δ transcript 3 (incorporating exon -2b) is only expressed in leukocytes as confirmed by large scale RT-PCR analysis

A more sensitive and high throughput PCR screen was designed in order to assess the presence of p110δ transcripts 2 and 3 (schematically shown in Figure 5-16) in different cell types.

This PCR made use of forward primers, specific for each of the 5' untranslated exons, in combination with a common reverse primer, complementary to a 20 bp sequence in exon 2 (note that in cells which express -2a or -2b, this PCR does not allow to screen specifically for the presence of transcripts containing only exon -1).

Total RNA from cells was extracted, converted to cDNA, and analysed with the 3 different sets of primers using the same PCR cycling program. The size of the PCR products was determined by agarose gel electrophoresis, and their sequence was verified by DNA sequencing (data not shown). We first analysed cells that differ most in their expression levels of p110δ mRNA and protein, namely mouse A20 lymphocytes *versus* NIH3T3 fibroblasts, and human U937 leukocytes *versus* HEK 293 epithelial cells (Chapter 3).

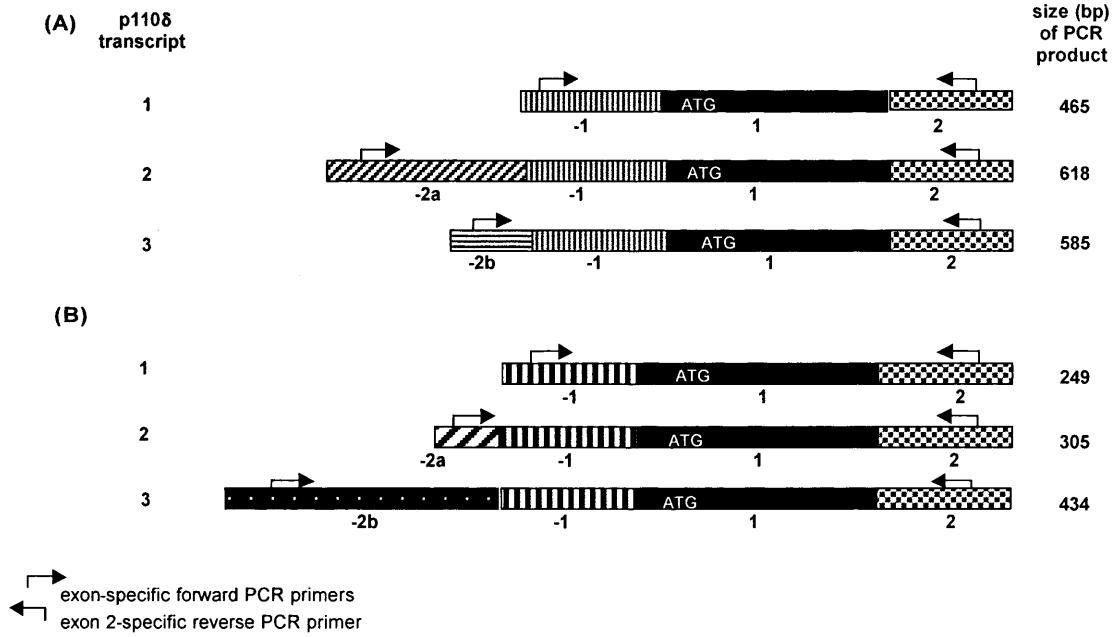

Figure 5-16 A RT PCR screen for the presence of distinct exons in the 5' UTR of the p110δ mRNA. A schematic representation of the position of the PCR primers used in the PCR screen to test expression of two of the distinct p110δ mRNA species in a panel of mouse **(A)** and human **(B)** cell lines. PCR using a forward primer in exon -1 identifies all 3 distinct transcripts.

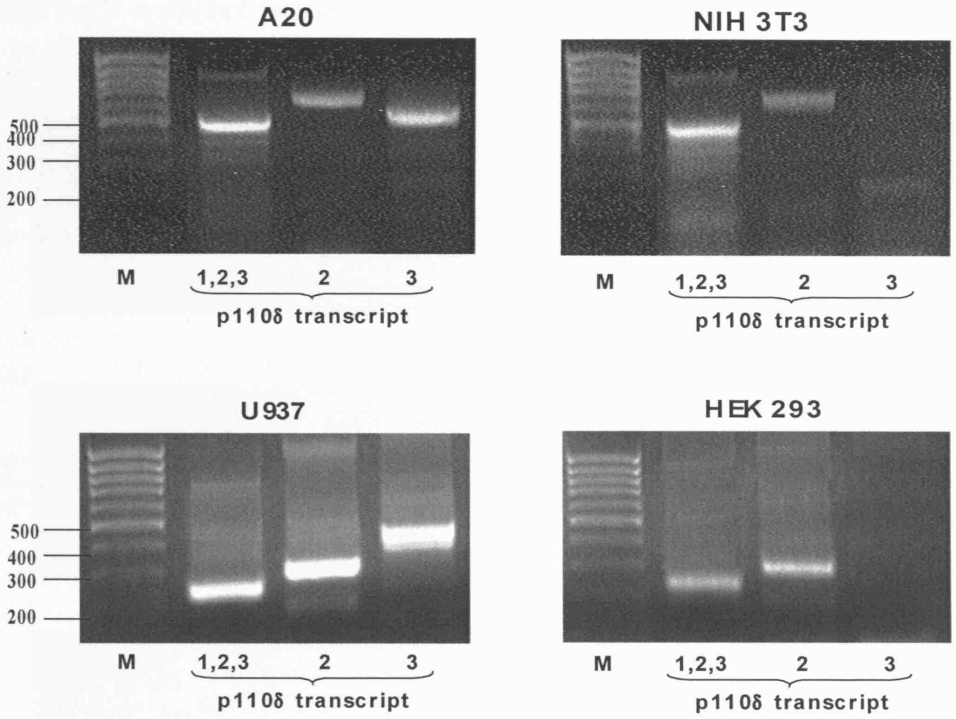

Figure 5-17 RT PCR screen for the presence of distinct exons in the 5' UTR of the p110δ mRNA. Detection of products of p110δ 5'-exon RT-PCR experiments by 2% agarose gel electrophoresis and ethidium bromide staining. n = 3

As can be seen from Figure 5-17 all cells were found to express transcript 2, which contains exon -2a and -1 (note that there was no evidence from the RACE analysis that any of these cell lines expressed transcripts containing exon -1 alone). In contrast, exon -2b transcripts were only found in the leukocyte cell lines, confirming our results obtained by RACE.

One clone, created during 5'RACE analysis of the mouse EL4 cell line identified an extra 5' untranslated exon, -2c. The RNAse protection probe designed to hybridise with exon -2c, failed to hybridise with a corresponding sequence in the mRNA of WEHI 231, EL4 or A20 cells. PCR primers were designed to amplify exon -2c to try and determine its authenticity (Figure 5-18). The PCR products were sequenced to assess the specificity of the PCR primers (data not shown).

The primers were designed with care yet failed to amplify a PCR fragment representative of exon -2c from total RNA extracted from six cell lines. As it is possible that the primers may not have worked it would have been desirable to test

them on the 5'RACE clone containing exon -2c. As this positive control was not carried out it is still not possible to conclude firmly that exon -2c is a cloning artifact.

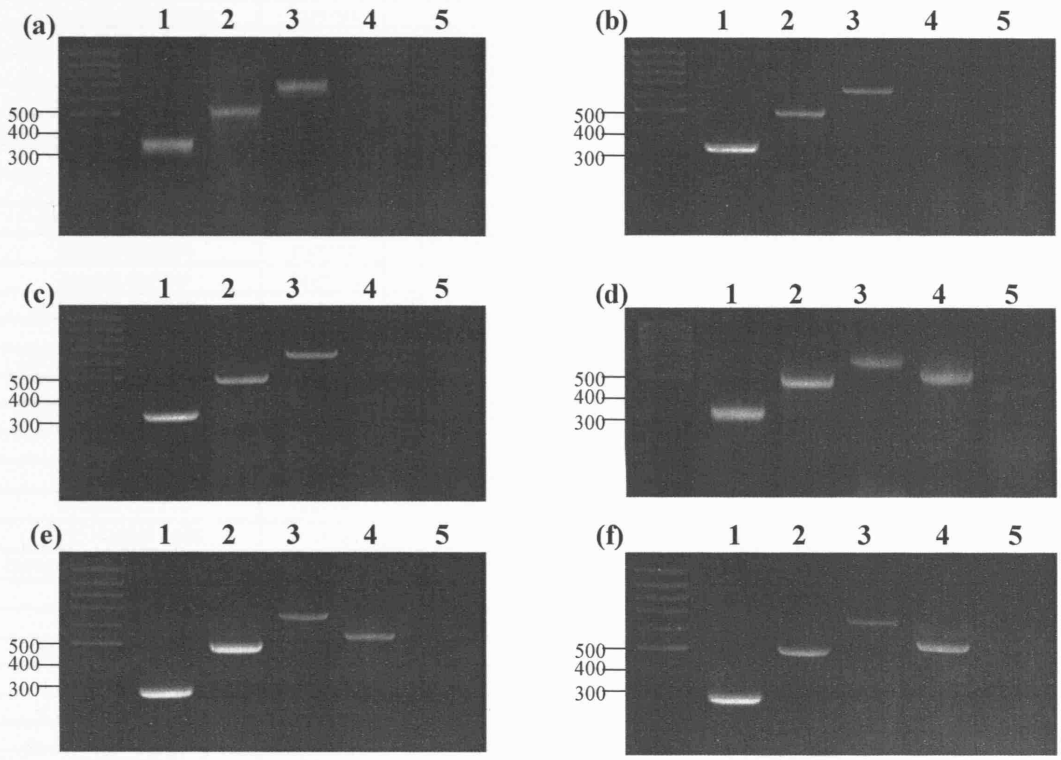

Figure 5-18 A RT-PCR screen of six mouse cell lines to corroborate the data on exon -2c obtained from RNase protection assay. RNA from three mouse non-leukocyte (a: NIH3T3, b: LLC and c: CT26) and three mouse leukocyte (d: A20, e: WEH 231 and f: EL4) cell lines was reverse transcribed and then subjected to multiple rounds of PCR amplification using exon-specific primers. The PCR products were resolved by 2% agarose gel electrophoresis and ethidium bromide staining. n = 2

A screen of a larger panel of cell lines (Table 5-4) corroborated the 5'RACE finding that only leukocyte cell lines express transcript 3 (incorporating exon -2b), in addition to transcript 2 (incorporating exon -2a). In our screen, non-leukocyte cell lines do not generate transcript 3, but instead generate a single p110δ mRNA, either transcript 1 or 2 (containing exon -1 or -2a, respectively).

Data from public EST databases support the notion that transcript 3 is mainly expressed in leukocytes (see Table 5-3 in section 5.2.3); note that when tissues are used in these experiments (such as aorta and vein, inner ear etc), it is difficult to exclude the presence of leukocytes). Taken together, the data above suggest that leukocytes can use alternate promoters to produce different p110δ mRNA species.

Cell type	Organism	Cell line	p110d transcript	exon -2b	exon -2a	exon -1	exon 1
breast malignant	human	MB-MDA 431	1			+	+
cervical carcinoma	human	HELA	1			+	+
breast cancer	human	PMC 42	1			+	+
breast cancer	human	MB-MDA134	1			+	+
breast cancer	human	CAMA1	1			+	+
breast cancer	human	T47D	1			+	+
breast cancer	human	GI101	1			+	+
breast cancer	human	MCF7	1			+	+
breast malignant	mouse	EMT 6	1			+	+
breast-benign	mouse	NMuMG	1			+	+
breast cancer	human	734B	2		+	+	+
breast cancer	human	MB-MDA 157	2		+	+	+
breast cancer	human	MB-MDA 231	2		+	+	+
breast cancer	human	HMT 3552	2		+	+	+
breast cancer	human	BT 20	2		+	+	+
breast cancer	human	MB-MDA 468	2		+	+	+
breast cancer	human	MB-MDA 175	2		+	+	+
breast cancer	human	HBL 100	2		+	+	+
breast cancer	human	MB-MDA 361	2		+	+	+
endothelial cells	human	EAHy 926	2		+	+	+
human embryonic kidney	human	HEK 293	2		+	+	+
epithelial	mouse	LLC	2		+	+	+
epithelial	mouse	CT26	2		+	+	+
embryonic fibroblast	mouse	MEF	2		+	+	+
fibroblast	mouse	NIH3T3	2		+	+	+
fibroblast	mouse	L929	2		+	+	+
melanoma	mouse	B16-F10	2		+	+	+
melanoma	mouse	melan-b	2		+	+	+
melanoma	mouse	B16-BL6	2		+	+	+
Acute monocytic	human	THP 1	2		+	+	+
leukaemia			3	+		+	+
monocyte	human	U937	2		+	+	+
			3	+		+	+
Tcell lymphoma	human	Jurkat	2		+	+	+
			3	+		+	+
Tcell leukaemia	human	CCRF-CEM	2		+	+	+
			3	+		+	+
T cell leukaemia	mouse	EL4	2		+	+	+
			3	+		+	+
B cell leukaemia	mouse	WEHI 231	2		+	+	+
			3	+		+	+
Pre B cell leukaemia	mouse	A20	2		+	+	+
			3	+		+	+

Table 5-4 The presence of different p110δ transcripts in a panel of human and mouse cell lines, as revealed by RT-PCR. n = 1

It is clear that mouse breast cells such as B16-BL6, melan-b and B16-F10 only express transcript 2 (Table 5-4). However from western blot analysis it is evident that the cells produce higher levels of p110δ protein than other non-leukocyte cell lines tested, (such as NIH 3T3; an example can be seen in Figure 3-3). The PCR screen above does not discriminate between high and low levels of p110δ mRNA. In order to

assess whether the high p110δ protein expression in some non-leukocyte cell lines relates to the relative levels of specific p110δ transcripts, we carried out a p110δ transcript specific real time PCR to assess the relative levels of the individual transcripts in specific mouse and human cell lines.

5.2.6 Relative expression of transcript 2 and transcript 3 in leukocyte and non-leukocyte cell lines

We wanted to know how certain non-leukocyte cell types, such as breast and melanoma) express high levels of p110δ protein without producing the –2b containing p110δ transcript. Real time PCR is a much more accurate method of determining the actual RNA levels of a particular gene than traditional PCR.

Results discussed in chapter 3 demonstrated a correlation between the levels of p110δ protein and mRNA expression levels in the mouse. The real time PCR primers used for the p110δ gene real time PCR analysis in chapter 3 were complementary to sequences in exons 8 and 9 (human) and exons 1 and 2 (mouse). These primers detect *all* p110δ transcripts and do not distinguish between the three different transcripts identified by 5'RACE.

Leukocytes express high levels of the p110δ protein and contain a p110δ mRNA transcript not present in other cell types. Melanoma cell lines, such as B16-BL6, also express relatively high level of p110δ protein but do not express transcript 3. An attempt was made to see if we could identify whether this extra transcript was responsible for the remarkable difference in the expression of the p110δ protein, and if so, whether was there a link between the high levels of p110δ protein and alternative transcript expression in the mouse melanoma cell lines.

Two sets of transcript-specific mouse and human p110δ primers were designed for real time PCR experiments. The primer set for transcript 2 (incorporating exon -2a) contained a forward primer complementary to exon -2a, a reverse primer complementary to exon -1 and an internal primer complementary to the 3' end of

exon -2a and the 5'end of exon -1, therefore it annealed across the exon junction. The primer set for transcript 3 (incorporating exon -2b) was similar to the set for transcript 2 except the forward primer was complementary to exon -2b and the internal primer annealed across the exon junction between exon -2b and exon -1 (Figure 5-19). Standard curve experiments were undertaken to ensure the efficiency of each primer set was compatible to enable us to compare data between primer sets.

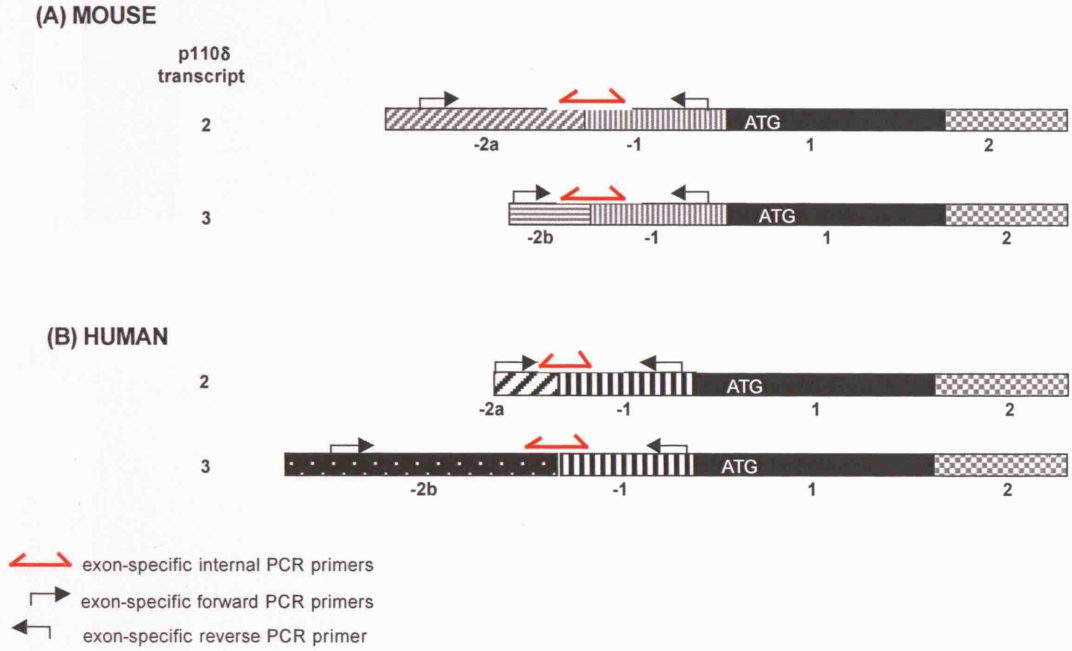

Figure 5-19 Schematic representation of the position of the PCR primers used in the real time PCR analysis of two of the major transcripts of the p110δ gene.

Using exon -2b-specific real time PCR primers in this set of experiments further corroborated our findings that transcript 3 is leukocyte-specific (yellow bars, Figure 5-20). Using these same primers there was no amplification of DNA from any of the non-leukocyte cell lines tested.

From the graphs in Figure 5-20 it is clear that the relative amount of transcript 2 (incorporating exon -2a, purple bars) is higher in leukocytes than in non-leukocyte cell lines suggesting that it is the combination of the expression of both transcripts that results in the elevated levels of p110δ protein in leukocyte cell lines and is not, as we had previously thought, due simply to the expression of transcript 3 (incorporating exon -2b).

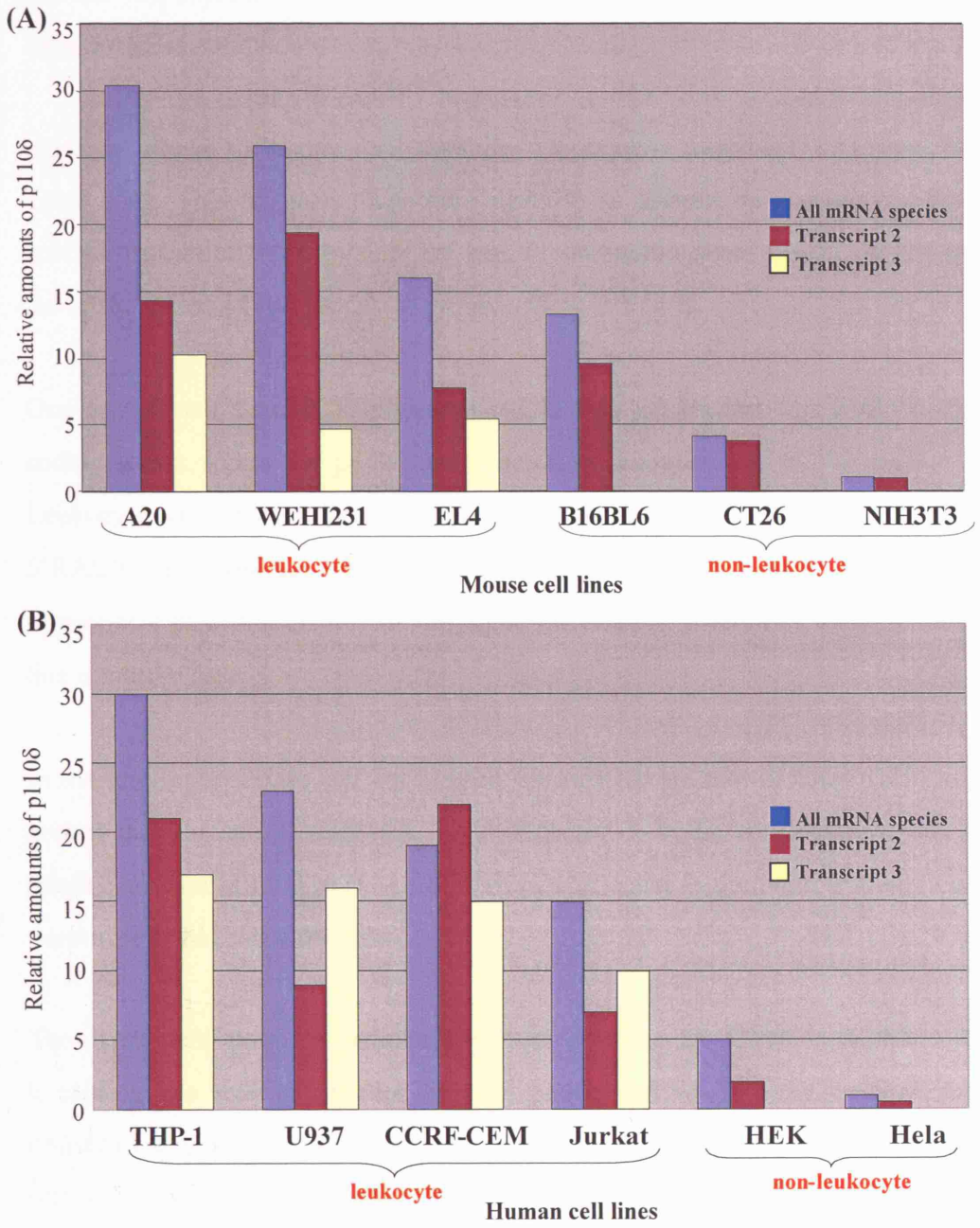

Figure 5-20 Relative amount of different p110δ mRNA transcripts in different cell lines (A) Transcript specific real time PCR results from three leukocyte and three non-leukocyte mouse cell lines. **(B)** Transcript specific real time PCR results from four leukocyte and two non-leukocyte human cell lines. n = 3

All leukocytes express transcript 3. There was no evidence that the higher p110δ protein expression in leukocytes is due solely to the expression of the unique leukocyte specific transcript, transcript 3. B16BL6, a mouse melanoma cell line (Figure 3-3), B16-F10 and melan-b, two other mouse melanoma cell lines tested (data

not shown), express p110δ protein at a higher level than the other non-leukocyte cell lines tested.

However, these melanoma cell lines do not express transcript 3 (Figure 5-20 and Table 5-4). This suggests that there may be an alternative mechanism, such as a specific transcription factor binding site, or chromatin deregulation, which results in the upregulating p110δ protein expression these cell lines.

One set of real time PCR primers used in this experiment are complementary to coding exons within the p110δ gene therefore recognize all p110δ mRNA species. Leukocyte cell lines express transcript 2 and transcript 3, as demonstrated by 5'RACE. Therefore the relative amount of p110δ mRNA recognized by primers for transcript 2 plus transcript 3 should equal the relative amount of total p110δ mRNA, this is not the case.

In the leukocyte mouse cell lines tested the relative amount of total p110δ mRNA is greater than the sum of transcript 2 and transcript 3. In the non-leukocyte cell lines the relative amount of total p110δ mRNA should be equal to the relative amount of p110δ transcript 2, this is not the case.

There are two possible reasons for these results. 1) There is a third transcript increasing the relative amount of total p110δ mRNA but not being detected by transcript 2 or transcript 3 specific real time PCR primers. 2) The efficiency of the 3 sets of primers is not exactly 100% therefore making such a direct comparison between the primers is not possible.

The sum of the relative value of p110δ transcript 2 and transcript 3 in the human leukocyte cell lines test was significantly higher than the relative value of total p110δ mRNA. The only explanation for these results is the difference in efficiency of the primers.

5.2.7 Database analysis of 5'UTRs belonging to other members of the Class I PI3K family and closely linked genes

The discovery of the three untranslated exons belonging to p110δ was unexpected. The phenomenon of untranslated exons is not new but the size of the intron separating the untranslated exons in p110δ is unusually large. To determine whether this detail was common among other genes a large scale analysis of a number of genes, human and mouse, was carried out using the Blast Like Alignment Tool (BLAT) (186). This analysis included the catalytic subunit of other members of the class I PI3K family, PIK3CA (p110α), PIK3CB (p110β) and PIK3CG (p110γ), the regulatory subunits of the classIA PI3Ks, PIK3r1 (p85α), PIK3r2 (p85β) and PIK3r3 (p55γ) (Figure 5-21). Other genes included in the analysis were Rho/RacGEF, PDGF, PDGFR, EGFR, and IL6. The results of the analysis did not identify any 5'untranslated exons in any of the non-PI3-kinase gene (data not shown). No 5' untranslated upstream exons were identified belonging to the PIK3r3 gene, unlike all the other members of the PI3K gene family analysed.

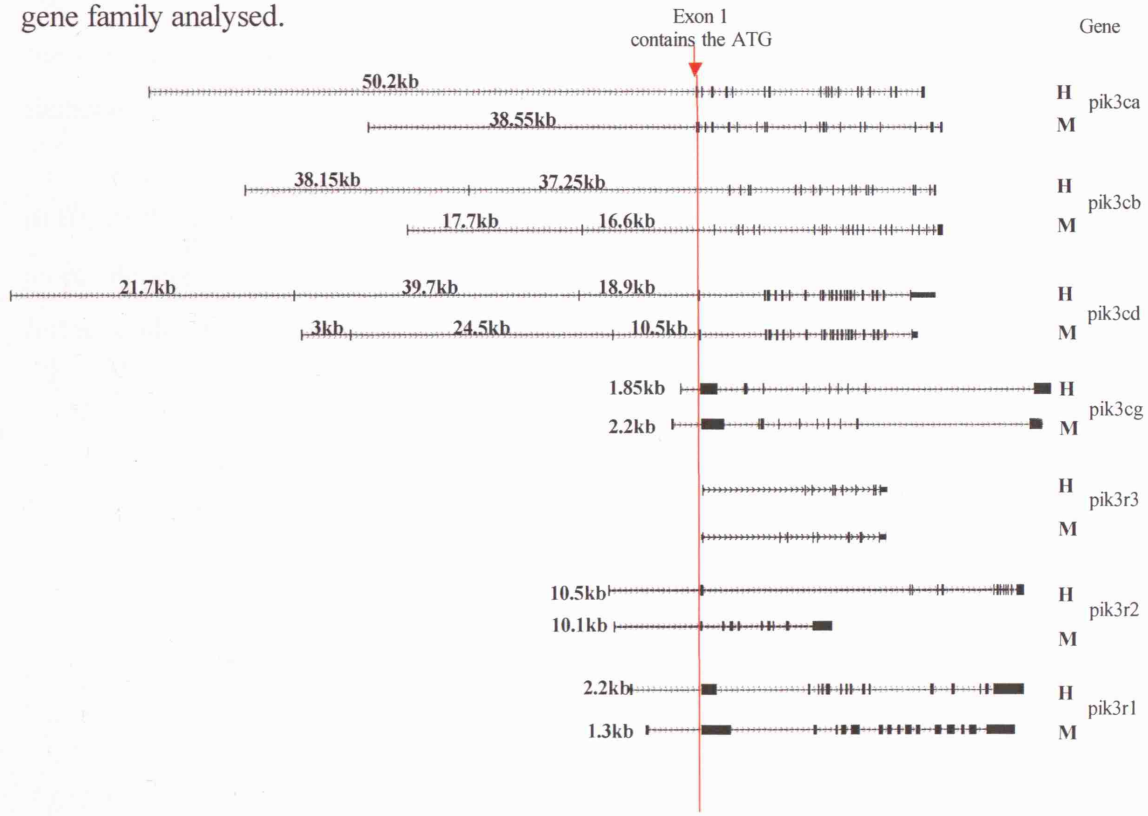

Figure 5-21 The combined results from a database analysis of the genomic profile of the other members of the Class I PI3Kinase family, the catalytic subunits and the class IA regulatory subunits. Solid bars correspond to the exons while the arrows along the thin lines represent the introns. The thin dotted line marks the beginning of the translated sequence. **H** and **M** indicate the human and mouse transcripts respectively. The size of each intron, belonging to the 5'UTR, is marked on the diagram. The picture is not to scale.

The regulation of gene expression of all of the catalytic subunits, and two out of the three regulatory subunits, of Class I PI3K appears to be quite complex. One untranslated exon was identified upstream of the mouse and human p110α gene, p110γ gene, p85α and p85β. The human and mouse p110β gene has two 5' untranslated exons. The p110δ gene, in human and mouse, is the only gene analysed that contains three 5' untranslated exons and it is the only class IA PI3K that exhibits tissue-specific gene expression. This evidence supports the notion that the expression of the extra 5' untranslated exon (exon -2b) in the p110δ gene may be the regulating factor that drives the tissue-specific expression of the gene.

p110γ exhibits tissue-specific gene expression but only has one 5' untranslated exon. This does not conform to the hypothesis stated above, that the expression of a third 5' untranslated exon may be the regulating factor that drives the tissue-specific expression of the gene. p110γ is a member of the Class IB PI3K genes and therefore may not be regulated in the same manner. It is also possible that more comprehensive analysis of the upstream region of the p110γ gene may identify further regulatory elements.

p110δ is the most comprehensively studied PI3K gene therefore it is possible that as more detailed analysis is undertaken of the other members of the PI3K gene family further undiscovered untranslated exons may be revealed.

6 *In silico* analysis of putative p110δ promoter elements

6.1 Introduction

A promoter is defined as a regulatory region a short distance upstream from the 5' end of a transcription start site that acts as the binding site for RNA polymerase. It is the region of DNA to which the RNA polymerase binds in order to initiate transcription. The identification of promoters and their regulatory elements is one of the major challenges in bioinformatics and integrates comparative, structural, and functional genomics. Many different approaches have been developed to detect conserved motifs in genes with homology between species. However, although recent approaches seem promising, in general, unambiguous identification of regulatory elements is not straightforward. The delineation of promoters is even harder, due to its complex nature, and *in silico* promoter prediction is still in its early years.

The aim of this part of my PhD project is to try to unravel the regulatory network controlling the expression of the p110δ gene. In this part of the project we tried to identify homologous areas between the mouse and human 5'untranslated region and characterise putative binding sites of known transcription factors. We also looked for known conserved regions in promoters such as the TATA box, CCAAT box and CpG islands.

The first step towards the identification of a transcription factor-binding site is finding the promoter. Despite the importance of promoters, our ability to identify these is not as good as our ability to identify coding regions. This is because promoters are very diverse, even well known motifs are not always conserved in all promoters. Furthermore each gene contains a set of unique transcription factor binding sites in the promoter that determines its temporal and spatial expression. The human genome contains about 1850 unique transcription factors (206), each of which binds to a specific transcription factor-binding site. The number of different combinations and permutations of transcription factor binding is immense.

The signal-based approach is the main approach taken to identify promoter regions. It relies on the recognition of relatively conserved signals such as the TATA box, CCAAT box and TFBS. Many bioinformatics programs used to identify promoter regions predict about 10-55% of the promoters correctly but at the same time produce a considerable number of false positives. First EF, a recently developed system, was reported to give considerably reduced levels of false positives. Many vertebrate promoter regions coincide with CpG islands. Therefore the signal-based approach combined with bioinformatics and with CpG island information, greatly enhances the accuracy of the analysis.

One major difficulty in identifying the regulatory network is due to the fact that the transcription of genes is not only regulated through the binding of the different transcription factors on the corresponding sites within the promoter region, but is also regulated through the structure of the DNA. This specific structure, which can change the accessibility of the binding sites, can greatly influence the expression of a gene. The structure of the DNA is regulated by the binding of histones and the forming of nucleosomes or by binding of other proteins that can alter the structure of the DNA.

6.2 Results

6.2.1 In silico analysis of the mouse and human p110δ gene with search parameters set at high stringency

6.2.1.1 Investigation into the homology between the coding region of the mouse and human p110δ gene

One possible way to identify regulatory elements involved in the expression of the p110δ gene is to compare the genome sequences of the p110δ gene in human and mouse. This is based on the assumption that the regulation of expression of the p110δ gene is similar between these two species. Whereas some evidence for this has been found it is important to note that, at present, we have no formal proof that the regulation of p110δ expression is the same between the two species. However the correlation between the expression of exon -2b-containing mRNAs and p110δ expression in leukocytes from both species is 100%, suggesting some functional basis for the comparative analysis described below.

LAGAN pairwise alignment software was used to identify regions of homology between the mouse and human p110δ gene. LAGAN is a pairwise global alignment program that allows the user to submit their own search parameters. The LAGAN server has been combined with the VISTA server to produce alignments and visualizations for all sequences submitted. Vista Browser is designed to visualize multiple large-scale alignments. The browser makes it easy to identify regions of high conservation across multiple species. The "peaks and valleys" on the graphs (see for example Figure 6-1) represent percent conservation between aligned sequences at a particular point on the base sequence. The top and bottom percentage bounds are shown to the right of every row.

The strong homology between the mouse and human p110δ coding sequence has been described before in Clayton *et al.* (204). The first sequences submitted to LAGAN were genomic DNA sequences of the mouse and human p110δ gene starting at the ATG translation start site and ending at the poly A tail. The parameters for this

alignment were set so that the homologous sequence had to be greater than 100 bp with over 75% homology (> 100 bp; >75%). The peaks in Figure 6-1 depict these regions.

The colouring of the different conserved regions corresponds to the annotation of the region. The pink regions are "Conserved Non-Coding Sequences" ("CNS"), the dark blue regions are conserved areas within the exons, and the light-blue regions are conserved areas in the 3'UTR of the p110δ gene. An arrow above the graph indicates the direction of the p110δ gene.

This LAGAN alignment demonstrated that every coding exon in the mouse p110δ gene shares over 75% homology with the corresponding exon in the human p110δ gene. Very little of the non-coding region between exon 1 and 22 is homologous between the mouse and the human p110δ gene.

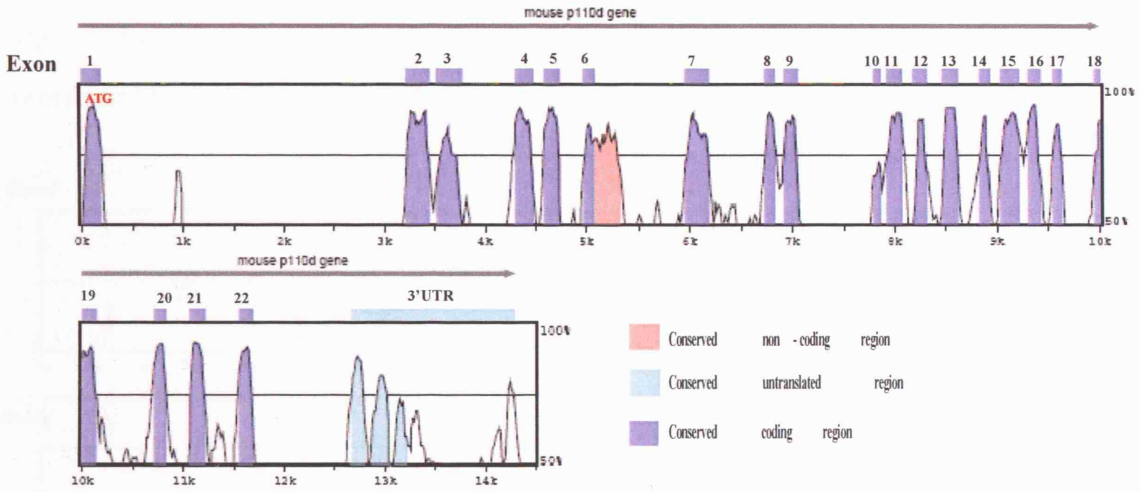

Figure 6-1 Visual representation (VISTA plot) of the conserved regions between the mouse and human p110δ gene identified by LAGAN pairwise alignment. This graph shows the percent of conservation between the genomic region from the translation start site up to the 3'UTR of the mouse and human p110δ gene. The stringency was set to > 100 bp region with over 75% homology. The top and bottom percentage bounds are shown to the right of the rows. The pink regions are "Conserved Non-Coding Sequences" ("CNS"), the dark blue regions are conserved regions within exons, and the light-blue regions are conserved regions within the 3'UTR. An arrow above the graph indicates the direction of the p110δ gene.

6.2.1.2 *Investigation into the homology between the 5' UTR of the mouse and human p110δ gene*

In an attempt to identify homologous regions within, and in the genomic region flanking, the newly identified 5' untranslated exons of the human and mouse p110δ gene, the corresponding sequences were submitted to LAGAN. Sequences from exon 1 (contains ATG translation start) up to and including 8 kb upstream of exon -2b in both mouse and human were submitted to LAGAN for analysis. Parameters were set at >100 bp with >75% homology. Seven regions of high homology were identified (A-G Figure 6-2). When the homologous sequences were analysed it became apparent that the alignment is not co-linear. The position of one of the regions (**A**-Figure 6-2), relative to the untranslated exons, was not conserved. For example, homologous region **A** lies upstream of exon -2b in the mouse p110δ gene (and is also represented like this in Figure 6-2) but upstream of exon -2a in the human p110δ gene, this is further schematically illustrated in Figure 6-3. The remaining six homologous regions identified were all downstream of exon -2a (Figure 6-3), ruling these out as putative promoters for transcripts 2 and 3.

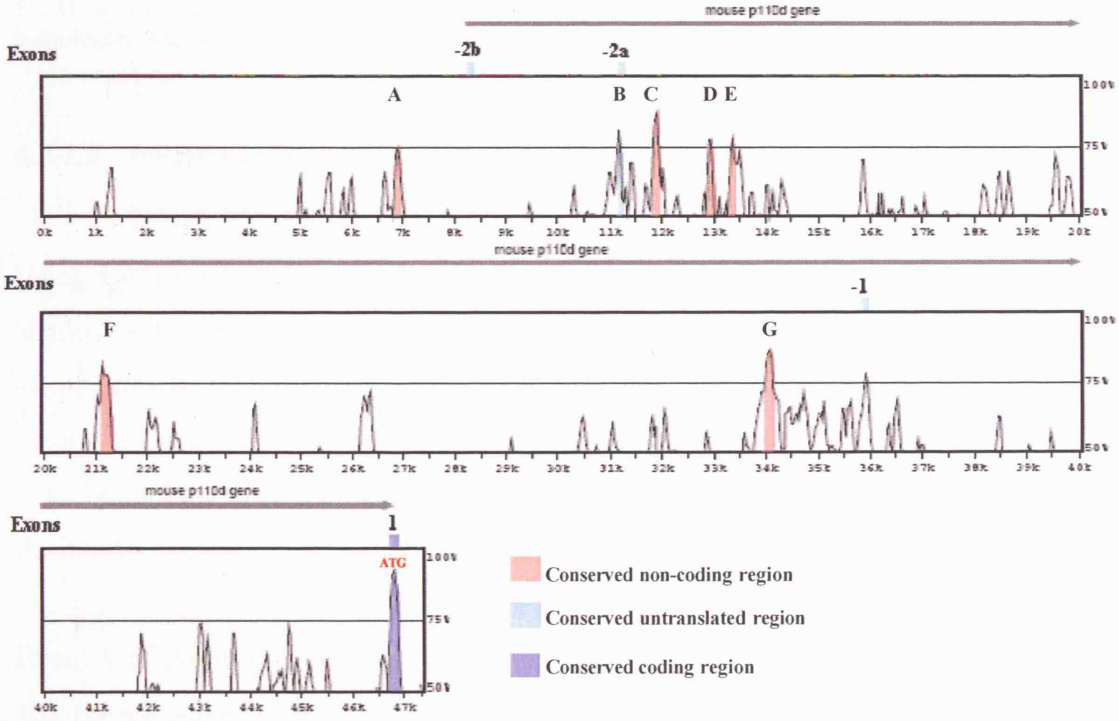

Figure 6-2 Visual representation (VISTA plot) of the conserved regions between the 5'UTR of the mouse and human p110δ gene, identified by LAGAN pairwise alignment. The stringency was set at > 100 bp region with a homology of over 75%.

All seven of these homologous regions were detected at a very high stringency level. Although they may not be promoters for the corresponding transcript, it is highly likely that they have been conserved for a regulatory reason, with enhancer or even repressor functions.

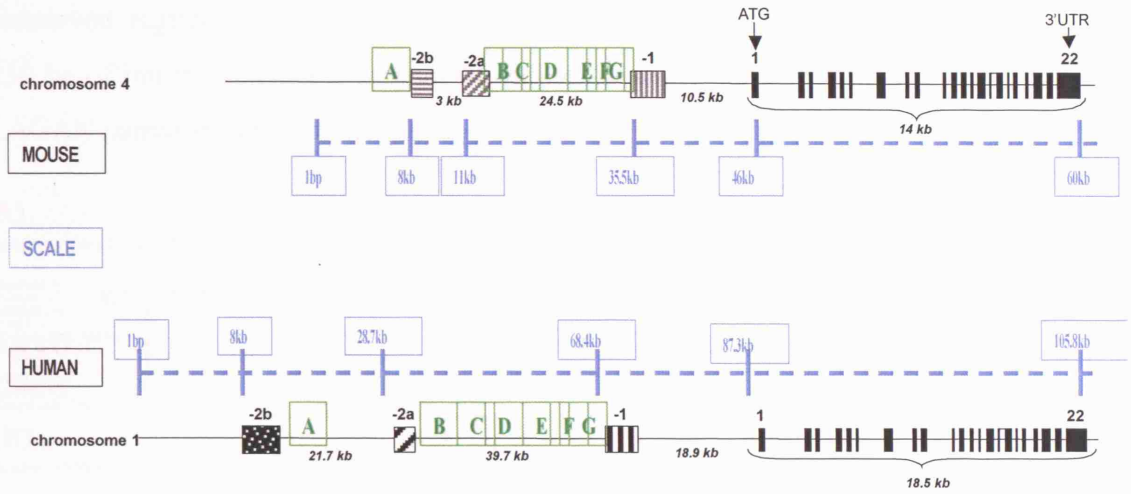

Figure 6-3 Genomic organisation of the mouse and human p110δ gene. In the red is a scale to illustrate the similarity in size in the coding region (exon 1-22) and the vast difference in the size of the 5'UTR between the mouse and human p110δ. The green lettering represents positions of the highly homologous regions identified by LAGAN pairwise alignment.

6.2.1.3 Investigation into the homology between the 5' untranslated exons of the mouse and human p110δ gene

Using LAGAN pairwise alignment exon -2a and -2b of the mouse p110δ gene were aligned with the corresponding exons in the human gene. No homology was identified, even at a low stringency level of > 5 bp with > 70% homology. This may be due to the considerable size difference between the exons. Indeed, mouse exon -2b is 80 bp and the corresponding human exon is 251 bp. Mouse exon -2a is 142 bp and the human exon is 72 bp.

Exon -1 of the mouse and human p110δ gene are of a more similar size (120 bp and 105 bp respectively) and have a significant sequence homology as can be seen from the sequence highlighted in gray in the alignment in Figure 6-5 A.

6.2.1.4 Investigation into the homology between the 5' untranslated exons, including 250 bp upstream of each exon, of the mouse and human p110δ gene

If exons are not conserved it is possible that the regions immediately upstream of the exons are conserved for regulatory purposes. Therefore in an attempt identify conserved region the three mouse and human p110δ-untranslated exons, including 250 bp of intron sequence upstream of each exon were aligned with each other using LAGAN pairwise alignment (Figure 6-4).

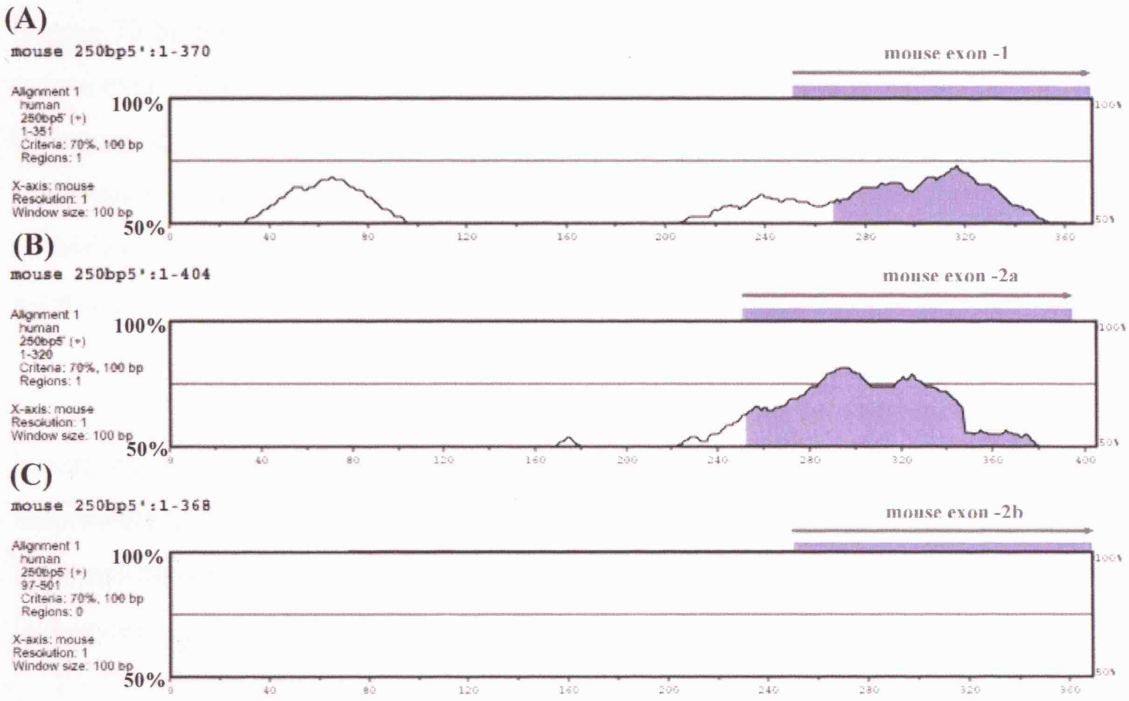

Figure 6-4 Visual representation (VISTA plot) of the conserved regions between the 5' untranslated exons in the mouse and human p110δ gene identified by LAGAN pairwise alignment. The parameters were > 100 bp with > 75% homology. Areas shaded in gray/blue are regions that LAGAN have identified as being conserved. (A) Mouse and human alignment of exon -1 including 250 bp of intron sequence 5' of the exon. (B) Mouse and human alignment of exon -2a including 250 bp of intron sequence 5' of the exon. (C) Mouse and human alignment of exon -2b including 250 bp of intron sequence 5' of the exon.

Sometimes the visual representation created by the LAGAN software can be misleading. It only allows the graph to be annotated with one of the aligned sequences, for example, either mouse or human. Therefore it does not depict which area of the human sequence has aligned. For example, the ~350 bp of sequence surrounding and including exon -1 a 98 bp region with 73.5% homology between the species was identified (gray/blue region Figure 6-4A). This is the same homologous region identified in section 6.2.1.3.

However, 138 bp of DNA was identified with 75.4% homology when the mouse and human exon -2a were aligned, including an extra 250 bp of DNA from the 5' intron sequence. From the 138 bp homologous region (grey/blue region in Figure 6-4B) 60 bp was human exon -2a sequence, and 78 bp was 5' intron sequence. This is not evident on the LAGAN visual representation. The alignment is shown more clearly in Figure 6-5B, here the mouse and human exon sequence is highlighted in grey.

I believe the similarity in size difference between the mouse and human exon -2a (mouse 70 bp bigger) and the length of human 5' intron sequence that aligns with the mouse exon -2a (78 bp) are too similar to be a coincidence. It is possible that the 78 bp human intron sequence that aligns with mouse exon -2a may once have belonged to human exon -2a but through evolution human may have lost the need for the 5' end of this exon. Alternatively we may have failed to identify the full length of the human exon -2a.

Using these very strong parameters no homology was found between the mouse and human exon -2b or in the 250 bp upstream of the exon (Figure 6-4C). This is unfortunate as we were especially interested in this area, as it could contain sequences important to drive expression of p110δ mRNA from exon –2b which is only found in leukocytes.

The actual DNA sequences of the above alignments are shown in Figure 6-5. The mouse p110δ gene sequence is shown on the top strand and the human sequence on the bottom strand. Figure 6-5A and B are the alignment between the mouse and human exon -1 and -2a, respectively.

(A)

```
>mouse 250bp5':1-370 (+)
>human 250bp5':1-351 (+)

000000001  G------------------------------TGTATACTACTTCTATGCTACCAACGGGA  000000030
>>>>>>>>>                                  |||       |||  | | |||| ||||  <<<<<<<<<
000000001  ACATTCCTGCTCCAGGCACTGGAAATGCTCCTGTGGTTGGTTTCAGTACCACCAAAGGGA  000000060

000000031  ACTGGGC-TGGGGTTGACACAGGTATCGAGAAAGGCCTACCCAGCAAAGTGGGGTTGTCT  000000089
>>>>>>>>>   |||||  |||  ||||| ||| |  | ||||||||| | | ||   | || | |||  |  <<<<<<<<<
000000061  GCTGGGGGTGGCATTGACGCAGATGGCAAGAAAGGCCGATCAAGTGGAATGTGATTGGTT  000000120

000000090  TACAAGAAGG-GAATCTGAGATGAAATGGCCCTTCCTTCAGTGCTGGGGGCTTTTTGCAT  000000148
>>>>>>>>>  | ||| |||| ||  |  |||||||   |    |  |                        <<<<<<<<<
000000121  TGCAAAAAGGAGAGACCTAGATGAAGCAGTTGATTTT-----------------------  000000157

000000149  GTTAGGCAGGCACGGCCACTGAGCCCTGAGATATATCATCAGCATGTACTGGATTCTGAA  000000208
>>>>>>>>>                      |  || |  |     |  || ||| || || |||||||  <<<<<<<<<
000000158  ----------------------GAGCTAACGTCCTGTAAGAGAATGCACAGGTTTCTGAA  000000195

000000209  ATTCTGTGTT-TAAAGCGTTCTTTCCTCTTTGCTTTCTTTCAGACATCTAAGGAGCTGAG  000000267
>>>>>>>>>  ||||| | || | ||       |||| |  || |||||||||      | ||        |  <<<<<<<<<
000000196  ATTCTCTCTTGTTAAAATAGTTTTCTTTCTTTCTTTCTTTCTTTTCCCCAACAGATAAGG  000000255

000000268  AGCCAGGCAGAAGTGGGATGAAGCCCGCTGATGCCAAAGTACCTTTAATCTCCCAGGCAG  000000327
>>>>>>>>>  || |||||    | |||||||  | |  |||| | |||| | |||||||| |||||| |  <<<<<<<<<
000000256  AGTCAGGCCAGGGCGGGATGACACTCATTGATTCTAAAGCATCTTTAATCTGCCAGGCGG  000000315

000000328  AGGGGCCTTGGCTGGTGGTCCTTCTTGGCCCATACCAAAGTGG  000000370
>>>>>>>>>  ||||| ||| ||||||   | |||||||  | || ||| |      <<<<<<<<<
000000316  AGGGGGCTTTGCTGGT---CTTTCTTGGACTATTCCAGA----  000000351
```

(B)

```
>mouse 250bp5':1-404 (+)
>human 250bp5':1-320 (+)

000000001  AGAGCTCCTAAGACAGTAGCCACTCACCCTGTAGTACTCTTAGCTTGTCTAGGTTTTGGT  000000060
>>>>>>>>>                                                               <<<<<<<<<
000000001  T-----------------------------------------------------------  000000001

000000061  TTTGCTTTGTTGTTGTTTTCTGAGTGACTGTGGCGATGCTCAGGGCTCTCCAGACCTTTA  000000120
>>>>>>>>>                                                               <<<<<<<<<
000000000  -----------------------------------------------------------  000000000

000000121  TTTGGCAGTGTGCCTTTGCATCAAGGGAAGACTTTGTTTTTGAAC---------AGAGCT  000000171
>>>>>>>>>             |||||||||  ||  ||| ||||  ||| |  |           ||     <<<<<<<<<
000000002  ------------CCTTTGCATAAAAAGAAAACTTGTTTTGTAGAAGGGCAAGCAAGGTGC  000000049

000000172  GGGGTTCGAAGGCACCTGCTGGTGCTACCCTCGCAGGGCCCGCTGGGCTTGGGGCGGGGC  000000231
>>>>>>>>>   |  | |  |||||||||  | || | |||    || |   |  |  || |||   | |  <<<<<<<<<
000000050  AGCTTCCAGAGGCACCTGG-GATGATGCCCCTCTAGCGGTAGGCGAACTGGGGAGTGTGG  000000108

000000232  TGGA------------GAGGCGTAGAGGATCAGTTGTTTAGCCTGTTATCTGTAGAAAGG  000000279
>>>>>>>>>    |               | | | ||||| | ||||||||||||||||    ||  |||||  <<<<<<<<<
000000109  GCGGCGGGGGAAGGCGAGGCCCTGGAGGACCTGTTGTTTAGCCTGTTGCAAGTGAAAAGG  000000168

000000280  AAACAAAGTGGGAAGTGGAGTGTGCGGACTGTCAGTAGGCGGGCTGTCCCGCTGCGCGCC  000000339
>>>>>>>>>  |||||||||||||||||||||||||  || | | |||||||  | ||||  |||||||  <<<<<<<<<
000000169  AAACAAAGTGGGAAGTGGAGTGTGCGGGTTGGCGGCAGGCGGGGCGCCCCG--GCGCGCC  000000226

000000340  CCGCCTCT---------GGCTCACTCGCGCCTAGCCTTGGGGCTGCCAGC----------  000000380
>>>>>>>>>  |||||||          |||||||||||||| |||   |  ||| | |||          <<<<<<<<<
000000227  CCGCCTCCCTCCCTCGAGGCTCACTCGCGCCCAGCGCAGTCGCTCCGAGCGGCCGCGAGC  000000286

000000381  --------TCCGCCGACCCAGCTGC--TGGACGT  000000404
>>>>>>>>>          | |||   ||||||||   ||  ||  <<<<<<<<<
000000287  AGAGCCGCCCAGCCCTGCCAGCTGCGCCGGGAGT  000000320
```

Figure 6-5 LAGAN pairwise sequence alignment between mouse and human p110δ exons -1 and -2a. (a) The alignment between mouse and human exon -1, including 250 bp of intron upstream **(b)** The alignment between mouse and human exon -2a, including 250 bp of intron upstream. The exons are highlighted in gray, and the conserved TATA boxes, identified by rVISTA, are highlighted in pink.

6.2.1.5 *rVISTA analysis of conserved promoter elements in the 250 bp region immediately upstream of each 5'untranslated exon in the mouse and human p110δ gene*

RNA polymerase II promoters are constructed from various combinations of three conserved promoter elements within 200 bp of the transcription start site; (1) TATA box (ATATAAA) about 25-30 bp upstream of the transcription start (+1) (2) CAAT box (GGCCAATCT) 3) GC box (GGGCGG) to which SP1, stimulatory protein 1, binds.

Certain observations have been made about promoter elements: (1) No element is essential (2) Individual elements may be repeated multiple times (present in many copies) (3) Only the TATA box is limited to one copy per promoter (4) Only the TATA box displays a conserved position relative to the transcriptional start site (-25 bp) (5) CAAT and GC sequences increase the frequency of initiation at particular promoters (6) CAAT and GC sequences are orientation-independent.

Taking all of these facts and observations into consideration the same sequences as above were analyzed for specific transcription factor binding sites including TATA box, CAAT and GC using rVISTA software available on the LAGAN website (described in section 2.11.3). A conserved TATA box was identified upstream of both exon -1 and exon -2a (Figure 6-5). No other conserved transcription factor binding sites belonging to the TATA box, GC, CAAT or TFIID were identified in these regions by the rVISTA software.

6.2.2 Analysis of putative regulatory regions in the mouse and human p110δ gene using lower stringency

One major limitation in computationally identifying functional transcription factor binding sites is that many transcription factors bind to short degenerate sequence motifs 6-12 bp in length. Such sequences occur very frequently in a genome, and experimentally it has been shown that only a very small fraction of these predicted transcription factor binding sites are functionally relevant. Transcription factor binding sites conserved between species increased the probability of their functionality. Therefore less stringent parameters were used to analyse these same

sequences. The parameters used in the following experiments were sequences of >5 bp with over 90% homology, unless otherwise stated. Below, we describe a number of different bioinformatic analysis tools that we used in an attempt to identify other possible regulatory regions upstream of each of the p110δ 5' untranslated exons.

6.2.2.1 Transcription Element Search System - TESS

TESS is a web tool for predicting transcription factor binding sites in DNA sequences. It uses information and statistical data from the TRANSFAC, IMD, and CBIL-GibbsMat database to identify binding sites. TRANSFAC represents the most comprehensive collection of transcription factor binding specificities. Three sets of transcription factor binding sites were taken into consideration: (1) Those thought to be important for the forming of the preinitiation complex e.g. TFIID and TBP – TATA Binding Protein. TFIID is a multi-component transcription factor that recognizes and binds to the promoter DNA. TBP is a subunit of TFIID that recognizes the TATA element (2) Possible sites for enhancer, Sp1, or repressor, WT1, binding. (3) Those that are binding sites for leukocyte-specific transcription factors e.g. Ikaros family of transcription factors and Nuclear Factor of activated T cells (NF-AT) family.

6.2.2.2 GRAIL and Softberry CpG Island analysis

CpG islands are regions where CpG sequences are present at significantly higher levels than is typical for the genome as a whole. CpG sequences are relatively rare unless there is selective pressure to keep them, for example, when they are involved in the regulation of gene expression. In vertebrates 60-90% of all CpG islands are methylated which leaves only a small part of the genome methylation-free. Many of the remaining non-methylated CpG sequences are found in CpG islands (~ 15% of all CpG sequences in human DNA). These are typically common near transcription start sites, and may be associated with promoter regions. Bird and his colleagues identified specific transcriptional repressors that recognise methyl-CpG, called MeCP1 and MeCP2 (35,36). MeCP2 contains a methyl CpG DNA-binding domain for direct chromatin structure alteration as well as a repressor domain for long-range repression

(36). Whereas unmethylated CpG islands are not affected by this type of repression and have a more relaxed nuclease sensitive chromatin structure.

CpG islands are commonly defined as regions of DNA of at least 200 bp in length with a G+C content above 50% (207). The GrailEXP (CpG island) software was available from the MRC – Rosalind Franklin Centre for Genomics Research Bioinformatics database.

6.2.2.3 First EF and FPROM putative promoter analysis software

FPROM from bioinformatics.net uses algorithms to predict potential transcription start positions by linear discriminate analysis (LDA), which combines characteristic functional motifs and oligonucleotide composition of these sites along with statistical data. FPROM uses files with selected transcription factor binding sites from current databases.

Traditional gene-finding programs treat the translation start site as the 5' boundary of the gene. Until the publication of First EF there were no computational tools to predict the non-coding first exons. Over 40% of human genes have completely non-coding first exons. The better-known computational promoter identification tools, such as PromoterInspector, give an average one false positive every several thousand base pairs and have a sensitivity of around 50%, which is unsatisfactory. First EF uses different discriminante functions designed to find potential first splice donor sites as well as CpG-related and non-CpG related promoter regions. Michael Zhang and his team demonstrated that First EF predicts 86% of first exons with 17% false positives. The accuracy of First EF has been confirmed by applying it to fully sequenced chromosomes (208).

Table 6-1 is a collaboration of all of the identified putative transcription factor binding sites within a 2 kb region upstream of each of the 5' untranslated exons in the mouse and human p110δ gene. A detailed description of the methods and sequences used to identify the transcription factor binding sites, and their relevance to this project, comes later in this chapter.

5' untranslated exon	Transcription factor binding site in mouse	Distance from exon (bp)	Transcription factor binding site in human	Distance from exon (bp)
-2b	**SP-1**	**-1954**	**SP-1**	**-1993**
	IK-1	**-1937**	**IK-1**	**-1975**
	PU-1	**-1916**	**PU-1**	**-1954**
	IRF-2	**-1883**	**IRF-2**	**-1953**
	c-Ets-2	**-1600**	**c-Ets-2**	**-1578**
	c-Myb	**-480**	**c-Myb**	**-523**
	TFIID	-190	NFAT-P	-206
	NFAT-P	-124	c-Ets-2	-163
	TBP	-101	TFIID	-138
	TFIID	-100	c-Ets-1	-107
			c-Ets-1	-105
			IK-1	-20
-2a	**c-Ets-2**	**-950**	**c-Ets-2**	**-1133**
	IK-1	**-682**	**IK-1**	**-831**
	Oct-1	**-272**	**Oct-1**	**-317**
	TBP	-143	TATA Box	-235
	TATA Box	-104	NFAT	-115
	TFIID	-101	IK-1	-84
-1	**8 x SP-1 sites**	**-2075 to-1537**	**8 x SP-1 sites**	**-2073 to -1565**
	IRF-1	**-1921**	**IRF-1**	**-1933**
	NFAT-1	**-1254**	**NFAT-1**	**-1292**
	IRF-1	**-1181**	**IRF-1**	**-1206**
	PU.1	**-810**	**PU.1**	**-833**
	c-myb	**-350**	**c-myb**	**-302**
	IK-1	-152	TATA Box	-49
	Oct-1	-69		
	TATA Box	-37		
	NFAT-P	-22		
	PU.1	-20		

Table 6-1 A summary of the identified transcription factor binding sites upstream of the mouse and human 5' untranslated exons of the p110δ gene. The rows highlighted in grey are the homologous transcription factor binding sites between the mouse and human p110δ gene. The rest are non-homologous transcriptions factor-binding sites identified in either the mouse or the human p110δ sequences analysed.

Shown below is a summary of the putative promoter regions and CpG islands identified in the 5'UTR of the mouse and human p110δ gene by First EF, FPROM, GRAIL and Softberry.

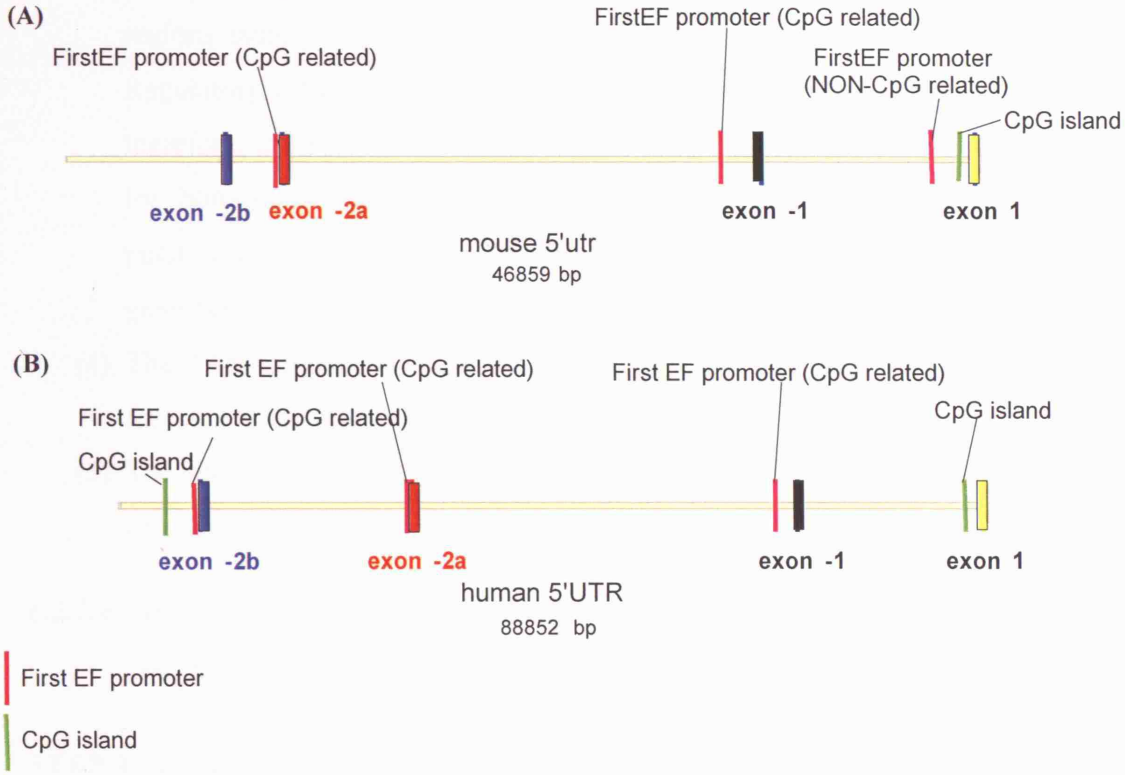

Figure 6-6 A diagrammatic representation of the promoters identified by the FIRST EF software and CpG islands identified by the GRAIL software in the 5'UTR of the mouse (A) and human (B) p110δ gene.

To construct the summary table and figure shown above, specific regions of the mouse and human p110δ 5'UTR were analysed using the different bioinformatic tools described previously. The upstream region of each exon (exon -2b, exon -2a, exon -1 and exon 1) was analysed in five stages:

(1) Starting with a small area of DNA, a 250 bp region upstream of the mouse and human exon was analysed for homology. Homologous regions identified in (1) were submitted to TESS to identify putative transcription factor binding sites involved in the regulation of p110δ gene expression.

(2) Although areas of homology are ideal places to start looking for putative regulatory regions, negative results do not rule out the possibility of there being area of DNA involved in regulation. They may be too far apart, in the mouse and human genome for the alignment software to recognise them as

homologous. Therefore the mouse and human p110δ sequences aligned in (1) were analysed individually using TESS.

(3) The failure to identify any homologous regulatory elements in the 250 bp regions studies in (1) and (2) we decided to expand our *in silico* studies. Regulatory elements can be many kb upstream of the transcription start site, therefore, a 2 kb region upstream of the mouse and human exon was analysed for homology. Homologous regions were submitted to TESS to identify putative transcription factor binding sites involved in the regulation of p110δ gene expression.

(4) The 2 kb sequence analysed in (3) was submitted to GRAIL and to Softberry for CpG island analysis

(5) The same sequence was also submitted to First EF and FPROM for putative promoter identification

6.2.2.4 In silico analysis of putative regulatory elements upstream of exon -2b in the human and mouse p110δ gene

STEP 1) LAGAN alignment of 250 bp of DNA immediately upstream of exon -2b in the mouse and human p110δ gene

Two regions of homology were identified in the intron sequence and three within the exon (shown in pink and blue, respectively Figure 6-7).

```
mouse  01  AGCCAGCTTTTTGTTTGTTTGTGGCTTTTTGGGTTTTTTTGTTTTTTGTTTTTGGTTTTT  000000060
>>>>>>>>>  ||  |  ||   |       |  |  |  |      ||   |      |  |  |  |     ||    ||||||   <<<<<<<<<
human  97  AGGCTGCGGTGCACGGGCTGGGGAGCACTTCTGCACCCTCGCTCTCCACTTCCTGTTTTTT  000000156
000000061  TTGAAACAAGGTGTTCTGGTTAGTTTTATGTCAACTTGACACAAGCTAGTGCCATCAGAG  000000120
>>>>>>>>>   |      | |    ||    |  | |   | | |||     |   |    |   ||||     | |||   <<<<<<<<<
000000157  CTCCCCAACGAATTTTGCGCCATTACTCTTTCATGGGGTGAGGGGGGAGTGGGGGCGGAG  000000216
000000121  AAGAGGGAGCCTCAATAGAGAAAATGCCTTTATAAAATCTGATTCTAGGCAAGCCTGCGG  000000180
>>>>>>>>>    | ||  ||||   | |  |  | |||         |   |    | |||  | |  | ||   <<<<<<<<<
000000217  --GCGGAGGCCTGCCTGGGAAGGAAGCCCGGGGGTCAGAGGCGCCCAGGACACCACGAGG  000000274
000000181  AGCATTGTCTCAATTAGTGATTGGTGTGGGAGTGC-------------------CCCGCC  000000221
>>>>>>>>>                    ||||    ||||||                       ||  ||   <<<<<<<<<
000000275  --------------------TTGGGAGAGGAGTGTGCTTTGCACTCTGCACTCTCCAGCT  000000314
000000222  CCTTGTGGGTGTGGCCTATCTTGAGCGGTAAGAAAGCAGGCCCAG---------------  000000266
>>>>>>>>>  |   || | || ||   ||  |  |  |   |   || ||||||                  <<<<<<<<<
000000315  CAGAGTAGCTGAGG---ATGCTAGGGAGACTGGGGGCTGGCCCACGGGGGGAGCAGAGGA  000000371
000000267  ----------------CCAGCCAGTCAGCAGCATTCCTCCGAGGTCTCTGCATCAACTCC  000000310
>>>>>>>>>                  ||  | |  | |||||||  |||   | ||    |    ||||   <<<<<<<<<
000000372  GTCCTCAGTGCGGTCCTCACTCTGAAAACAGCATTTTTCCCTAGGCTTGAGAATGCCTCC  000000431
000000311  TGCCCTGTGTGAGT---------------GTCTGTCCTGACTTCCTAAGAAGACGAACAG  000000355
>>>>>>>>>   |   ||      |               |||||||  |   | | | |   |     |   <<<<<<<<<
000000432  GGGATTGGACTGATCCAGGCTGGACCCACGTCTGTCTGGTGATACCAGGGGCAGAGGGAC  000000491
000000356  TGATGTAGAAGTG  000000368
>>>>>>>>>    | |  |     <<<<<<<<<
000000492  CACTCTGACA---  000000501
```

Figure 6-7 LAGAN alignment of mouse and human exon -2b including 250 bp of upstream intron sequence. The stringency was set at a minimum of 5 bp with over 90% homology. The top strand is mouse and the bottom strand is human. The region highlighted in pink depicts the homologous regions in intron sequences. Blue highlighted areas depict the homologous regions in exon sequence.

The two homologous sequences identified in the intron, GTTTTT and GGAGTG, highlighted in pink (Figure 6-7) were submitted to TESS for identification of possible transcription factor binding sites. None were identified.

STEP 2) TESS analysis of 250 bp of DNA immediately upstream of exon -2b in the mouse and human p110δ gene

TFIID binding sites were identified in both the mouse and the human 250 bp regions upstream of exon -2b. This is a transcription factor that recognizes and binds to promoter DNA binding sites. As mentioned earlier in the chapter the TATA box displays a conserved position relative to the transcriptional start. A component of TFIID is TBP, the TATA binding protein. The long distance of the TFIID binding sites from both the mouse and the human exons suggests that they are likely to be non-functional binding sites.

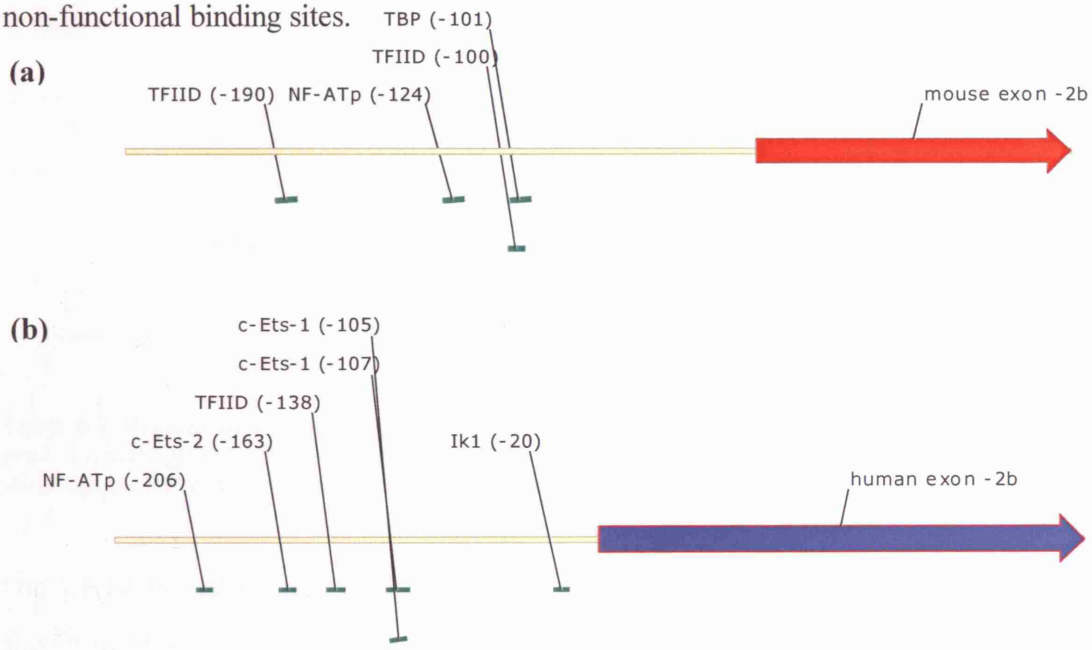

Figure 6-8 Vector NTI schematics of the positions of particular transcription factor binding sites upstream of mouse (a) and human (b) exon -2b.

An NF-ATp binding site was also identified in the same region. Two other leukocyte specific transcription factor-binding sites, c-Ets-1 and Ik1, were also identified in the human sequence but not in the mouse. The Ets family of genes encodes a class of transcription factors. Ets1 is predominantly expressed in the lymphoid organs of neonatal and adult mice, whereas Ets2 is expressed in every organ examined (209). The Ikaros gene family is a group of transcription factors expressed in haemopoietic

stem cells. Gene regulation studies have shown them to be involved in lymphocyte commitment and differentiation (210).

STEP 3) LAGAN and TESS analysis of a 2kb region DNA immediately upstream of exon -2b in the mouse and human p110δ gene

Homologous regions, identified using LAGAN, were submitted to TESS for transcription factor binding site identification. Table 6-2 shows the homologous regions that contain transcription factor binding sites of interest, their distance of the homologous area from exon -2b and the position and sequence of the identified transcription factor binding sites:

Homologous region	Homologous sequence	Distance from exon -2b (bp)		Putative transcription factor binding sites
		mouse	human	
1	GGGTGGTGGTGGCGCACAATCCCAGCACTCAGGAGG CAGAGGAA	1954	1993	Sp1, Ik-1 PU-1
2	CCAGCCTGGTCTACAAAGTGAG	1883	1953	IRF-2
5	TTTCCTTGCTCACAGGG	1600	1578	c-Ets-2
12	GTGAAGTTCAG	480	523	c-Myb

Table 6-2 Regions of homology in 2 kb upstream of exon -2b in the mouse and human p110δ gene. Included in the table are the distance of the homologous regions from the exon and potential transcription factor binding sites within these regions.

Out of 14 homologous regions identified (their positions, relative to exon -2b, are shown in Figure 7-3) 4 contained transcription factor-binding sites of putative targets for the regulation of the p110δ gene. Sp1 is a transcription factor whose binding leads to super-activation of the gene to which it has bound. Sp1 has been shown to synergistically activate a native human promoter in a cellular context. The effect of Sp1 depends on its distance from the TATA box. As no TATA box has been identified in this region it is likely that this is a non-functional the binding site. Ik-1 is a transcription factor involved in lymphocyte commitment and differentiation, therefore may be relevant for p110δ expression in leukocyte cell lines. PU1 is macrophage and B-cell specific but it is not expressed in T cells, which rules it out as a sole activator of p110δ gene expression.

Interferon regulatory factor 1 (IRF-1) is a transcriptional activator, and its antagonistic repressor is IRF-2. IRF-2-deficient mice exhibited bone marrow suppression of hematopoiesis and B lymphopoiesis and mortality suggesting a role in the immune system. c-myb is required for blood cell development *in vivo* and *in vitro*. However, very little is known about its mechanism of action.

The regions of homology where TESS did not identify any transcription factor binding sites does not rule them out as possible regulatory regions. Further analysis of these regions was done using reporter assays (Chapter 7).

STEP 4) GRAIL and Softberry CpG island analysis of 2 kb of DNA immediately upstream of exon -2b in the mouse and human p110δ gene

Using the GrailEXP and Softberry CpG island identification software further *in silico* analysis was done on the 2 kb region upstream, and including exon -2b in the mouse and human p110δ gene. No CpG islands were identified upstream of the mouse exon -2b by either program. GrailEXP identified a CpG island of 348 bp, 1435 bp upstream of human exon -2b with a GC content of 74.61%.

STEP 5) First EF and FPROM putative promoter analysis of 2 kb of DNA immediately upstream of exon -2b in the mouse and human p110δ gene

Using First EF and FPROM, two different pieces of promoter identification software freely available on the Internet, the upstream region of exon -2b in mouse and human were scanned for putative promoter regions. First EF and FPROM identified one putative promoter region. It covered 483 bp immediately upstream of the exon (Figure 6-9). The software classed it as a CpG related promoter. Strangely neither GrailEXP nor Softberry identified this region as CpG rich.

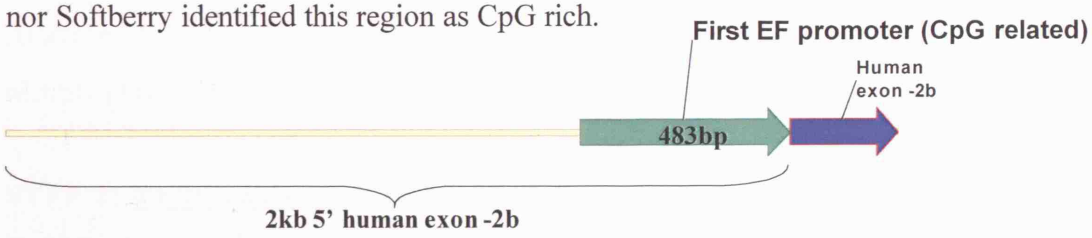

Figure 6-9 Schematic representation of the position of the putative promoter region upstream of the human p110δ exon -2b.

6.2.2.5 *Putative regulatory elements upstream of exon -2a in the human and mouse p110δ gene*

STEP 1) LAGAN alignment of 250 bp of DNA immediately upstream of exon -2a in the mouse and human p110δ gene

Three regions of homology were identified in the intron sequence (shown in pink, Figure 6-10). Six regions of homology within the exon were also identified (shown in blue Figure 6-10).

```
mouse  01  AGAGCTCCTAAGACAGTAGCCACTCACCCTGTAGTACTCTTAGCTTGTCTAGGTTTTGGT  000000060
>>>>>>>>>                                                               <<<<<<<<<
human  01  T-----------------------------------------------------------  000000001
000000061  TTTGCTTTGTTGTTGTTTTCTGAGTGACTGTGGCGATGCTCAGGGCTCTCCAGACCTTTA  000000120
>>>>>>>>>                                                               <<<<<<<<<
000000000  -----------------------------------------------------------  000000000
000000121  TTTGGCAGTGTGCCTTTGCATCAAGGGAAGACTTTGTTTTTGAAC---------AGAGCT  000000171
>>>>>>>>>             |||||||||  ||  |||  ||||  |||  |  |          ||    <<<<<<<<<
000000002  ------------CCTTTGCATAAAAAGAAAACTTGTTTTGTAGAAGGGCAAGCAAGGTGC  000000049
000000172  GGGGTTCGAAGGCACCTGCTGGTGCTACCCTCGCAGGGCCCGCTGGGCTTGGGGCGGGGC  000000231
>>>>>>>>>   |  | |  |||||||||  | ||  |  |||    ||  |   |  |  ||  |||   | |   <<<<<<<<<
000000050  AGCTTCCAGAGGCACCTGG-GATGATGCCCCTCTAGCGGTAGGCGAACTGGGGAGTGTGG  000000108
000000232  TGGA------------GAGGCGTAGAGGATCAGTTGTTTTTCCTGTTATCTGTAGAAAGG  000000279
>>>>>>>>>     |               | | |  |||||  | ||||||||||||||||    ||  |||||  <<<<<<<<<
000000109  GCGGCGGGGGAAGGCGAGGCCCTGGAGGACCTGTTGTTTTTCCTGTTGCAAGTGAAAAGG  000000168
000000280  AAACAAAGTGGGAAGTGGAGTGTGCGGACTGTCAGTAGGCGGGCTGTCCCGCTGCGCGCC  000000339
>>>>>>>>>  |||||||||||||||||||||||||||  ||  |  |  |||||||  |  ||||  |||||||  <<<<<<<<<
000000169  AAACAAAGTGGGAAGTGGAGTGTGCGGGTTGGCGGCAGGCGGGGCGCCCCG--GCGCGCC  000000226
000000340  CCGCCTCT---------GGCTCACTCGCGCCTAGCCTTGGGGCTGCCAGC----------  000000380
>>>>>>>>>  |||||||          |||||||||||||| |||   |  |||  |  |||          <<<<<<<<<
000000227  CCGCCTCCCTCCCTCGAGGCTCACTCGCGCCCAGCGCAGTCGCTCCGAGCGGCCGCGAGC  000000286
000000381  --------TCCGCCGACCCAGCTGC--TGGACCG  000000404
>>>>>>>>>           | |||   ||||||||   ||  ||  <<<<<<<<<
000000287  AGAGCCGCCCAGCCCTGCCAGCTGCGCCGGGACG  000000320
```

Figure 6-10 LAGAN alignment of mouse and human exon -2a including 250 bp of upstream intron sequence. The stringency was set at > 5 bp with > 90% homology. The top strand is mouse and the bottom strand is human. Pink highlighted region depict the homologous regions in intron sequence. Blue highlighted regions depict the homologous regions in exon sequence.

The three homologous sequences identified in the intron, CCTTTGCAT, AGGCACCTG and GAGGA (Figure 6-10), were submitted to TESS for identification of possible transcription factor binding sites. None were identified.

STEP 2) TESS analysis of 250 bp of DNA immediately upstream of exon -2a in the mouse and human p110δ gene

The 250 bp mouse and human p110δ sequences aligned in step 1 were analysed individually using TESS. Transcription factor binding sites of interest were identified (Figure 6-11).

A TATA box was identified, but it is too far away from the transcription start site of both the mouse and the human exon -2a to be credible. However the leukocyte-specific, NF-AT and IK-1, transcription factor binding sites may well be functional. Although exon -2a is expressed in both leukocyte and non-leukocyte cell lines, it appears to be expressed at a much higher level in leukocyte cell lines (as described in chapter 5.2.6). Therefore it may be the use of these transcription factors that enables the leukocytes to do this.

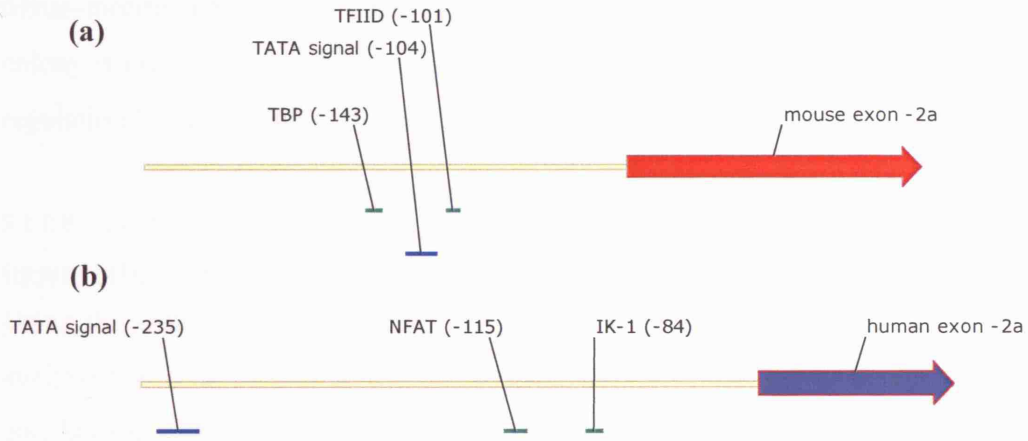

Figure 6-11 Vector NTI schematics of the positions of particular transcription factor binding sites in a 250 bp area upstream of mouse (a) and human (b) exon -2a.

STEP 3) LAGAN and TESS analysis of a 2kb region DNA immediately upstream of exon -2a in the mouse and human p110δ gene

Further analysis of the 2 kb region upstream of exon -2a in the human and mouse p110δ gene was undertaken using LAGAN alignment. Homologous regions that were identified in this way were then submitted to TESS for transcription factor binding site identification. A table of the homologous regions, their distance from the exon and any identified binding sites are shown below:

Homologous region	Homologous sequence	Distance from exon -2a (bp)		Putative TF binding sites
		mouse	human	
2	AAGGAATAC	950	1133	c-Ets-2
5	AGTGCTGGGATTA	682	831	Ik-1
10	CCTTTGCATCAA	272	317	Oct-1

Table 6-3 Regions of homology in 2 kb upstream of exon -2a in the mouse and human p110δ gene. Included in the table are the distance of the homologous regions from the exon and potential transcription factor binding sites within these regions.

Of the 11 homologous regions identified (their positions, relative to exon 1, are shown in figure 7-7) 3 contained transcription factor-binding sites that could potentially be involved in the regulation of the p110δ gene.

c-Ets-2 and Ik-1 have been described above. The Oct-1 transcription factor regulates a variety of tissue-specific and general housekeeping genes by recruiting specialized co-activators of transcription. It is ubiquitously expressed but has been seen to exert tissue-specific control on genes such as interleukins, the granulocyte-macrophagal colony-stimulating factor and immunoglobulin α. It exerts its tissue-specific regulation by interacting with tissue-specific co-activators (211).

STEP 4) GRAIL and Softberry CpG island analysis of 2 kb of DNA immediately upstream of exon –2a in the mouse and human p110δ gene

Using the GrailEXP and Softberry CpG island identification software further *in silico* analysis was done on the 2 kb region upstream, and including exon -2a in the mouse and human p110δ gene. GrailEXP and Softberry searches identified a CpG island immediately upstream of the mouse and human exon -2a that was 264 bp and 255 bp with a GC content of 65.15% and 70.3% respectively.

STEP 5) First EF and FPROM putative promoter analysis of 2 kb of DNA immediately upstream of exon -2a in the mouse and human p110δ gene

The same 2 kb region analysed above was submitted to First EF and FPROM. Both programs identified a CpG-related promoter immediately upstream of exon -2a in both the mouse and human p110δ gene. This data correlates nicely with the CpG island predictions described in STEP 4 above.

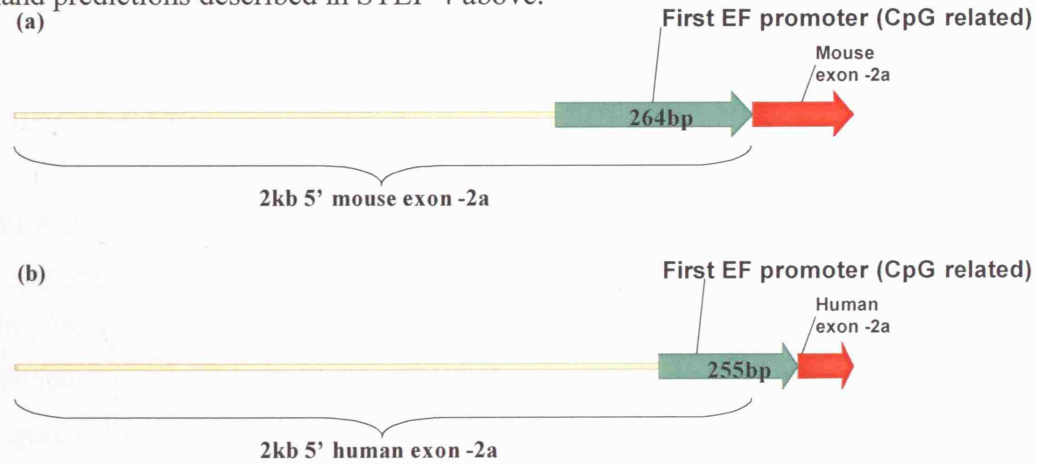

Figure 6-12 Schematic representation of the position of the putative promoter region upstream of the human p110δ exon -2a.

6.2.2.6 *Putative regulatory elements upstream of exon -1 in the human and mouse p110δ gene*

STEP 1) LAGAN alignment of 250 bp of DNA immediately upstream of exon -1 in the mouse and human p110δ gene

Eight regions of homology were identified in the intron sequence (shown in pink). Seven regions of homology in the exon were also identified (shown in blue) (Figure 6-13).

```
Mouse 001  G------------------------------TGTATACTACTTCTATGCTACCAACGGGA  000000030
>>>>>>>>>                                 |||       |||  | | ||||| ||||  <<<<<<<<<
Human 001  ACATTCCTGCTCCAGGCACTGGAAATGCTCCTGTGGTTGGTTTCAGTACCACCAAAGGGA  000000060

000000031  ACTGGGC-TGGGGTTGACACAGGTATCGAGAAAGGCCTACCCAGCAAAGTGGGGTTGTCT  000000089
>>>>>>>>>   |||||  |||  ||||| ||| |  | |||||||||| | | ||   | || | |||  |  <<<<<<<<<
000000061  GCTGGGGGTGGCATTGACGCAGATGGCAAGAAAGGCCGATCAAGTGGAATGTGATTGGTT  000000120

000000090  TACAAGAAGG-GAATCTGAGATGAAATGGCCCTTCCTTCAGTGCTGGGGGCTTTTTGCAT  000000148
>>>>>>>>>  | ||| |||| ||  |  |||||||   |    |  |                        <<<<<<<<<
000000121  TGCAAAAAGGAGAGACCTAGATGAAGCAGTTGATTTT-----------------------  000000157

000000149  GTTAGGCAGGCACGGCCACTGAGCCCTGAGATATATCATCAGCATGTACTGGATTCTGAA  000000208
>>>>>>>>>                     |  || |  |     |  || ||| || || |||||||  <<<<<<<<<
000000158  ----------------------GAGCTAACGTCCTGTAAGAGAATGCACAGGTTTCTGAA  000000195

000000209  ATTCTGTGTT-TAAAGCGTTCTTTCCTCTTTGCTTTCTTTCAGACATCTAAGGAGCTGAG  000000267
>>>>>>>>>  ||||| | || | ||      |||| |  || |||||||||      | ||        |  <<<<<<<<<
000000196  ATTCTCTCTTGTTAAAATAGTTTTCTTTCTTTCTTTCTTTCTTTTCCCCAACAGATAAGG  000000255

000000268  AGCCAGGCAGAAGTGGGATGAAGCCCGCTGATGCCAAAGTACCTTTAATCTCCCAGGCAG  000000327
>>>>>>>>>  || |||||    | |||||||  | |  |||| | |||| | ||||||||| |||||| |  <<<<<<<<<
000000256  AGTCAGGCCAGGGCGGGATGACACTCATTGATTCTAAAGCATCTTTAATCTGCCAGGCGG  000000315

000000328  AGGGGCCTTGGCTGGTGGTCCTTCTTGGCCCATACCAAACAGG  000000370
>>>>>>>>>  ||||| ||| ||||||   | ||||||| | || ||| |      <<<<<<<<<
000000316  AGGGGGCTTTGCTGGT---CTTTCTTGGACTATTCCAGA----  000000351
```

Figure 6-13 LAGAN alignment of mouse and human exon -1 including 250 bp of upstream intron sequence. The stringency was set at a minimum of 5 bp with over 90% homology. The top strand is mouse and the bottom strand is human. The region highlighted in pink depicts the homologous regions in intron sequences. Blue highlighted areas depict the homologous regions in exon sequence.

The eight homologous sequences identified in the intron, highlighted in pink (Figure 6-13) were submitted to TESS for identification of possible transcription factor binding sites. None were identified.

STEP 2) TESS analysis of 250 bp of DNA immediately upstream of exon -1 in the mouse and human p110δ gene

The mouse and human p110δ sequences aligned in Figure 6-13 were analysed individually using TESS. Transcription factor binding sites of interest were identified (Figure 6-14).

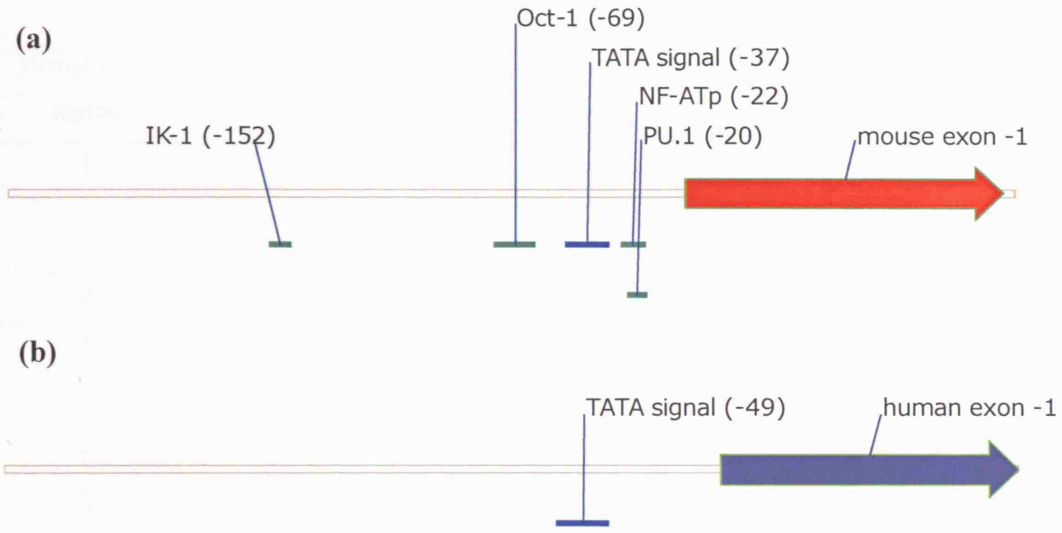

Figure 6-14 Vector NTI schematics of the positions of particular transcription factor binding sites upstream of mouse (a) and human (b) exon -1.

The search highlighted a TATA signal at positions -37 and -49, upstream of exon -1 in the mouse and human p110δ gene respectively (Figure 6-14). This element was also identified by rVista (Figure 6-5). These TATA elements are close enough to the conserved position of a TATA box (~-25 and –30) for them to be considered as potential regulatory elements of the p110δ gene. Other leukocyte-specific transcription factors, Ik-1, Oct-1, NF-ATp and PU.1, were identified upstream of the mouse exon -1 but not the human.

The roles of Ik-1, Oct-1 and NFAT have been described above. PU.1 plays an important role in haematopoiesis. Mice that are deficient for the Ets family member PU.1 exhibit a profound defect in the generation of B, T and myeloid progenitors. Although PU.1 is known to be required for early lymphoid differentiation, how it actually regulates these developmental processes is unknown (212).

STEP 3) LAGAN and TESS analysis of a 2kb region DNA immediately upstream of exon -1 in the mouse and human p110δ gene

Further analysis of the 2 kb region upstream of exon -1 in the human and mouse p110δ gene was undertaken using LAGAN alignment. The 5 bp 90% stringency search identified over 75 regions of homology. The stringency was therefore set to a minimum of 10 bp with 90% homology. The homologous regions were submitted to TESS for transcription factor binding site identification. A table of the homologous regions, their distance from the exon and any identified binding sites is shown below:

Homologous region	Homologous sequence	Distance from exon -1 (bp)		Putative transcription factor binding sites
		mouse	human	
1	CGCCTCGCCCC	2075	2073	Sp-1
2	CAGCCCCGCCCCGAGCCGCGCCCC TCCGCCGA	2036	2047	Sp-1
3	GGGCCTGCCCTCAGGCCC	1975	1988	Sp-1
5	TCCGGTCACTTCGGGACCCTCCCG CCGTGTCCCCCGGGCAACTGTCTC CAGGGAGACGG	1921	1933	IRF-1, Sp-1
6	CCCGCCCC	1860	1871	Sp-1
8	CCCACCTGTGCGGGAGTCTG	1829	1840	Sp-1
10	GGGTGGACAGGGT	1735	1743	Sp-1
12	CTGCCGCCCTTGATG	1537	1565	Sp-1
18	ATTTTTTCCTGGCTCCTTATCTCC	1254	1292	NF-AT-1
21	GCTTTGTCTTTCC	1181	1206	IRF-1
27	TTTGTCTCATGGAGAAGCCAGT	810	833	PU.1
33	ACCAACGGGAACTGGG	350	302	c-myb

Table 6-4 Regions of homology in 2 kb upstream of exon -1 in the mouse and human p110δ gene. Included in the table are the distance of the homologous regions from the exon and potential transcription factor binding sites within these regions.

Out of 36 homologous regions identified (their positions, relative to exon -1, are shown in figure 7-1) 8 contained SP-1 transcription factor-binding sites that could possibly be involved in the regulation of the expression of p110δ transcript 1 (Table 6-4). Sp-1 has been shown to interact synergistically to activate transcription. Three Sp1-binding sites when present in the same promoter activate transcription to a level that is 8 times greater than would be expected given their individual activities in the absence of the other two sites (213). Within the 2 kb region upstream of exon -1 TESS identified 8 Sp-1 transcription factor binding sites. These sites may or may not be accessible to the Sp-1 transcription factor in the p110δ gene.

The roles of all of the other leukocyte-specific transcription factors highlighted in Table 6-4 have been mentioned above. Transcript 1 (incorporating exon -1 alone) is not leukocyte-specific therefore it is highly unlikely that these leukocyte-specific transcription factor-binding sites are functional. However, it is possible that they may

be involved in co-operative binding with transcription factors in other parts of the genome to increase transcription from an alternative transcriptional start site.

STEP 4) GRAIL and softberry CpG island analysis of a 2 kb region of DNA immediately upstream of exon -1 in the mouse and human p110δ gene

Using the GrailEXP and Softberry CpG island identification software further *in silico* analysis was done on the 2kb region upstream, and including exon -1 in the mouse and human p110δ gene. GrailEXP and Softberry searches identified a CpG island 2 kb and 1.9 kb upstream of the mouse and human exon -1 respectively. The mouse CpG island is 426 bp with a GC content of 71.78%. The human CpG island is 1038 bp with GC contents of 73.81%.

STEP 5) First EF and FPROM putative promoter analysis of a 2 kb region of DNA immediately upstream of exon -1 in the mouse and human p110δ gene

The same 2 kb region analysed above was submitted to First EF and FPROM. Both programs identified a CpG related promoter 2 kb upstream of mouse exon -1 and 1.9 kb upstream of human exon -1. This data correlates nicely with the CpG island predictions described above in step 4.

6.2.2.7 Putative regulatory elements upstream of exon 1 in the human and mouse p110δ gene

Vanhaesebroeck *et al* identified the ATG translation start site of the human p110δ gene (8), it was surrounded by an optimal Kozak consensus sequence and had an in-frame stop codon 12 nucleotides upstream. 5' RACE had also been performed on the mouse and human p110δ gene (188). The published sequences from these 5' RACE experiments did not identify the full length of either the mouse or the human p110δ gene. Therefore before I identified the 5' untranslated exons of the p110δ gene I carried out reporter assays on putative promoter elements within a 2 kb region upstream of exon 1, where I had originally believed the transcriptional start site to be.

Significant reporter activity was seen from some of the putative promoter elements (Figure 7-15).

STEP 1) LAGAN alignement of 250 bp of DNA immediately upstream of exon 1 in the mouse and human p110δ gene.

The elements with the greatest promoter activity were within 300 bp upstream of the exon. Therefore 475 bp and 473 bp upstream and including exon 1 of the mouse and human p110δ gene, respectively, were aligned using LAGAN pairwise alignment software. Five regions of homology were identified in the intron sequence (shown in pink in (Figure 6-15), none of which contained transcription factor binding sites. Homologous regions between mouse and human exon 1 are shown in blue (Figure 6-15).

```
Mouse 001  CTCT--GGCCAGAACATTTTTAAGAACTGAAGAGTTTTATCTTTGGTTTTCAAGCATCTA  000000058
>>>>>>>>>   ||   |    | |  ||  | || |  || || ||                         <<<<<<<<<
Human 001  TTCGTGGATGTGTATGTTCCTGAGGAAAGATGATTTG----------------------  000000037

000000059  CCTGTTCTTGTTTCCCACTGTGGGTCACCTGTCGGCCACAGTGCGGTCTTTTGGCCTATG  000000118
>>>>>>>>>   |||| | || | |    ||||||                              |   |||  <<<<<<<<<
000000038  -CTGTGCGTGCTCC----TGTGGGG--------------------------AGGACATG  000000065

000000119  GCATGTCCACCTGTAATAAGAAGCACCTCGGCCTATAGGGTGGCTTAAATTGGTGGGAGT  000000178
>>>>>>>>>     ||||  ||||  | |||| ||| ||  | |      |||||    | | || ||||  <<<<<<<<<
000000066  CAGTGTCTGCCTGGGAGAAGAGGCAGCTGAGGCC-----GTGGCCCGGAGTAGTAGGAGC  000000120

000000179  CACAAGCCACC-TAGGCAGGATGCTGCCTCATGTGTAGCAAGCTTTGCATGTCCTCCTGT  000000237
>>>>>>>>>  |||||||||||| | ||||||| |  |||    | |  |  ||||||   |  || |  ||  <<<<<<<<<
000000121  CACAAGCCACCCTGGGCAGGACGAGGCCCTGAGGGAGGTGAGCTTTTTGTACCCGCAGGT  000000180

000000238  GG----CCCAGATCTGGCCGTCCCATTGGCCAAAGCCT----------------------  000000271
>>>>>>>>>   |    | ||   |||  | |||   || |||| | |                         <<<<<<<<<
000000181  CGGGAACTCACTCCTGAGCTTCCTGCTGTCCAAGGACCCCACTGTTTTCCAGGGAGTCCC  000000240
000000272  ----------------------------TAGTTCATCATGCCTTTGCCTTTCCCACAG  000000301
>>>>>>>>>                             ||  ||||  |||| || | |       ||  <<<<<<<<<
000000241  TTCCAAAGGTCTCACCCAGCTCAGCTGAGGTAACTCATTTTGCCATTTCTTCATTTTTAG  000000300


000000302  GAAAACAGACCTCAGAGAAGTAACACAACAGGATGCCCCCTGGGGTGGACTGCCCCATGG  000000361
>>>>>>>>>  || ||| | | || | |||||||||   |||||||||||||||||||||||||||||||  <<<<<<<<<
000000301  GACAACTGTCATCTGGGAAGTAACAACGCAGGATGCCCCCTGGGGTGGACTGCCCCATGG  000000360

000000362  AGTTCTGGACCAAAGAGGAGAGCCAGAGCGTGGTTGTTGACTTCTTGCTGCCCACAGGGG  000000421
>>>>>>>>>  | ||||||||||| |||||||  ||||||||  || |||||||| ||||||||||||||  <<<<<<<<<
000000361  AATTCTGGACCAAGGAGGAGAATCAGAGCGTTGTGGTTGACTTCCTGCTGCCCACAGGGG  000000420

000000422  TCTACTTGAACTTCCCCGTGTCCCGCAATGCCAACCTCAGCACCATCAAGCAGG  000000475
>>>>>>>>>  |||||  |||||||||| ||||||||||||||||||||||||||||||||||   <<<<<<<<<
000000421  TCTACCTGAACTTCCCTGTGTCCCGCAATGCCAACCTCAGCACCATCAAGCAG-  000000473
```

Figure 6-15 LAGAN alignment of mouse and human exon 1 including 250 bp of upstream intron sequence. The stringency was set at a minimum of 5 bp with over 90% homology. The top strand is mouse and the bottom strand is human. The region highlighted in pink depicts the homologous regions in intron sequences. Blue highlighted areas depict the homologous regions in exon sequence.

STEP 2) TESS analysis of 250 bp of DNA immediately upstream of exon 1 in the mouse and human p110δ gene

The mouse and human p110δ sequences aligned in Figure 6-15 were analysed individually using TESS. Transcription factor binding sites were identified (Figure 6-16).

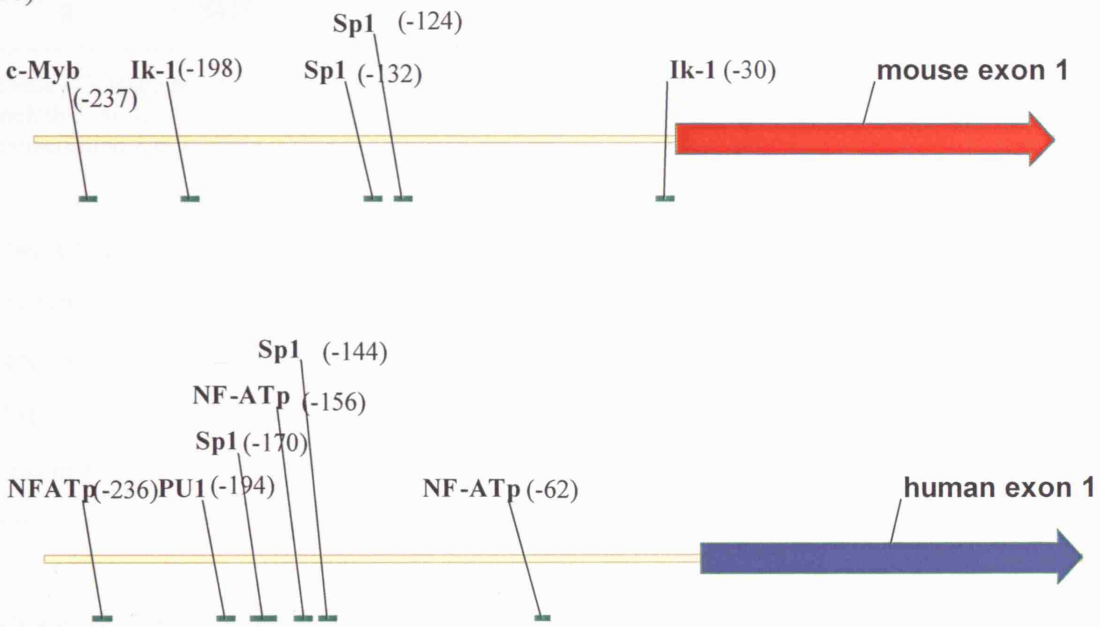

Figure 6-16 Vector NTI schematics of the positions of particular transcription factor-binding sites upstream of mouse (a) and human (b) exon 1.

Leukocyte-specific transcription factor binding sites are marked on Figure 6-16 for speculative purposes. There has been no evidence to suggest that any region upstream of exon 1 of the p110δ gene is involved in the tissue-specific regulation of the gene. Sp1 was the only non-tissue-specific transcription factor-binding site identified in this region in both the mouse and the human p110δ gene.

STEP 3) LAGAN and TESS analysis of a 2kb region DNA immediately upstream of exon -2a in the mouse and human p110δ gene

Further analysis of a 2 kb region upstream of exon 1 in the human and mouse p110δ gene was undertaken using LAGAN alignment. The homologous regions were submitted to TESS for transcription factor binding site identification. A table of the homologous regions, their distance from the exon and any identified binding sites are shown below:

Homologous region	Homologous sequence	Distance from exon -1 (bp)		Putative transcription factor binding sites
		mouse	human	
1	CTTTAATCCCAGC	1451	1872	Ik-2
3	CCTGGCCTACATAGTAGAGACTCTATTTCAAAA	1400	1791	TFIID
8	AAGAACTGAAG	455	564	c-Myb

Table 6-5 Regions of homology in 2 kb upstream of exon 1 in the mouse and human p110δ gene. Included in the table are the distance of the homologous regions from the exon and potential transcription factor binding sites within these regions.

Out of 10 homologous regions identified (their positions, relative to exon 1, are shown in figure 7-14) 3 contained transcription factor-binding sites that were possible targets for the regulation of the p110δ gene. A TFIID binding site was found but its distance from the 5' end of exon 1 suggests that it is non-functional. We also have no reason to believe that there is a functional promoter immediately upstream of exon 1. The possible function of the other transcription factors has been described above.

STEP 4) GRAIL and Softberry CpG island analysis of a 2 kb region of DNA immediately upstream of exon 1 in the mouse and human p110δ gene

Although there was no evidence from the 5'RACE analysis that transcription of the p110δ gene begins at exon 1, CpG islands tend to be unmethylated and are therefore not affected by methylation specific transcriptional repression. Their more relaxed chromatin structure may therefore allow transcription regulation via binding of enhancer or repressor proteins. Using the GrailEXP and Softberry CpG island identification software a 2 kb region upstream of, and including, exon 1 in the mouse and human p110δ gene was analysed. Softberry searches identified a CpG island 600 bp and 1.8 kb upstream of the mouse and human exon 1 respectively. The mouse CpG island is 198 bp with a GC content of 85.64%. The human CpG island is 199 bp with GC contents of 76%.

STEP 5) First EF and FPROM putative promoter analysis of a 2 kb region of DNA immediately upstream of exon 1 in the mouse and human p110δ gene

The same 2 kb region analysed above was submitted to First EF and FPROM. First EF identified a non-CpG related promoter 2 kb upstream of mouse exon 1 but no putative promoter region was identified upstream of the human p110δ exon 1. The accuracy for predicting non-CpG-related first exons is low due the statistical

composition of these exons is very similar to that of introns (208). The results from this promoter search help to clarify the results from the 5'RACE experiments that it is unlikely that there is a transcriptional promoter upstream of exon 1 in either the mouse or human p110δ gene.

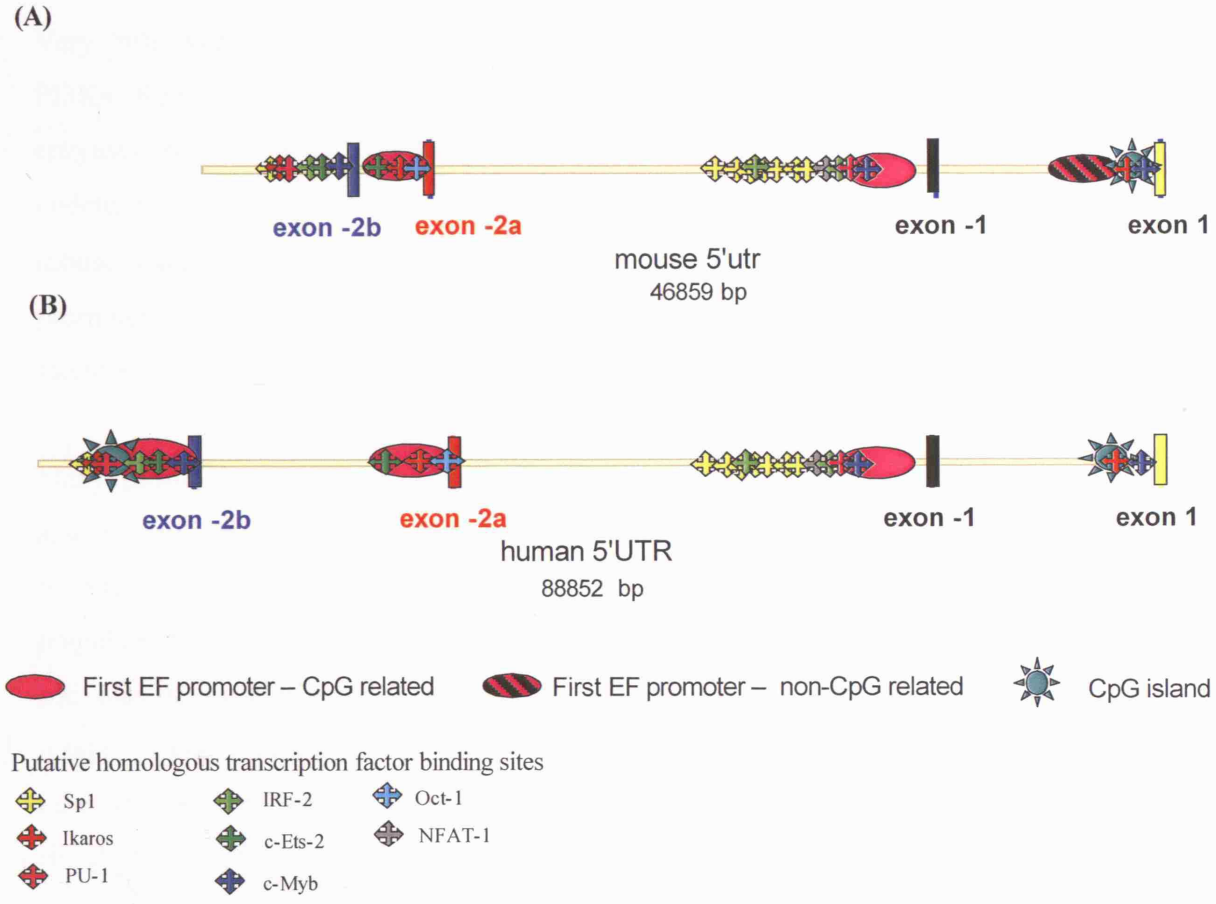

Figure 6-17 A summary model of the positions of the promoters identified by the FIRST EF software, CpG islands identified by the GRAIL software and putative transcription factor binding sites identified using TESS analysis in the 5'UTR of the (A) mouse and (B) human p110δ gene.

Bioinformatic analysis of sequences is a very good starting point for further *in vitro* analysis. Results obtained from bioinformatic studies are not conclusive. Information collated from this group of experiments was used as a guide in designing the reporter constructs for the more detailed *in vitro* analysis using reporter assays, described in Chapter 7.

7 p110δ promoter analysis using reporter assays

7.1 Introduction

Very little work has been carried out on the regulation of the expression of class I PI3Ks. Reports have been published detailing the pathways in which they are key enzymes, but the actual mechanism by which their expression is regulated is still undetermined. A region 418 bp in length, upstream of the transcription start site of the mouse class IB p110γ catalytic subunit, was reported as having tissue-specific promoter activity in a human leukocyte and non-leukocyte cell line, identified using a luciferase reporter assay (174).

Many promoter regions have been identified using reporter assays. Bioluminescent assays are widely used for screening cellular activities. The strategy of one particular reporter assay is to couple the Firefly luciferase gene to a specific DNA element that potentially regulates gene expression. In the dual luciferase assay system, the activities of the firefly and *Renilla* luciferase are measured sequentially from a single lysate. Upon completing the measurement of the firefly luciferase activity (the experimental reporter), the firefly luminescence is rapidly quenched, with simultaneous activation of the *Renilla* luciferase luminescent reaction (the control reporter activity). Thus, the system allows rapid quantification of both reporters co-expressed in the lysates of the transfected cells. This eliminates problems caused by different transfection efficiencies between samples and slight experimental errors such as the calculation of the number of cells successfully transfected and the number of viable cells post transfection.

Unfortunately there is no set formula with transfections. Different methods work for different cell lines, and developing an efficient and reproducible transfection protocol for each experimental system is not straightforward.

There is a choice of two types of transfection: stable transfection (where foreign DNA is integrated into the chromosomal DNA of the cell and maintained throughout the life of the cell line) and transient transfection (where DNA is introduced into the

nucleus but is not integrated into the chromosomes). Generating stable cell lines takes considerably longer but the DNA is maintained for the lifetime of the cell line.

By allowing more copies of the gene to be present in the nucleus, transiently transfected cells typically generate higher protein expression levels than stable cell lines. The quick turnaround time also makes transient transfections ideal for initial experiments.

There are only two types of transfection but a large number of different methods with which to transfect the cells. The artificial liposome was the first method used by us to deliver DNA into cells. Most lipid-based transfection reagents contain a mixture of positively charged lipids and neutral lipids. The positive end of the lipid molecule associates with the negatively charged DNA backbone, forming a liposome/nucleic acid complex. The positively charged complex associates with the cell membrane and is taken up by endocytosis. However, this method did not prove favorable with either NIH3T3 mouse fibroblasts or the A20 B cell lymphoma.

The Superfect reagent, from Qiagen corp., was chosen as the transfection method of choice for the mouse fibroblast cell lines, as it showed the greatest efficiency and the highest transfection efficiency for these particular cells. A20 mouse leukocytes however, failed to take up the reporter vectors using the Superfect method or the lipid-based transfection method. Therefore the more invasive method of electroporation of DNA was used. Electroporation-induced transfection is based on the agitation of the cell membrane by an electrical pulse, which forms temporary pores in the membrane and allows the passage of nucleic acid into the cell. Each cell type requires optimization of duration and strength of the electric pulse to provide a balance between the conditions that allow efficient DNA delivery and those that kill cells. Even optimized electroporation procedures result in 50% or greater cytotoxicity. It is clear that electroporation is a much more temperamental procedure than the other methods described above.

7.2 Results

7.2.1 Investigation of potential tissue-specific promoter elements upstream of exon -2b of the p110δ gene

Data obtained to date suggest that exon -2b, the more distal untranslated exon 5' from the ATG translation start site, is only expressed in leukocyte cell lines (as detailed in Table 5-4). This is true in both mouse and human cell lines.

LAGAN alignment software used to search for homologous sequences between the mouse and the human p110δ gene does not search in a co-linear fashion. This results in the identification of homologous sequences in different positions in the genomic sequence of the mouse and human gene. For example, a 100 bp sequence *upstream* of exon -2b in human aligns with sequence of similar size *downstream* of exon -2b in mouse.

Alignment of specific sequences can be forced by aligning smaller regions of DNA. For example if a region upstream of exon -2b in mouse is aligned with a region in the corresponding area in the human gene this prevents the software scanning the entire gene for homologous regions. Therefore smaller areas of the mouse and human p110δ gene were aligned (Section 6.2.2). It is well known that enhancer and suppressor control elements can exist at sites tens of thousands of bases upstream or even downstream of the transcription start site (TSS). In most cases however, the essential control elements are present within the proximal promoter a few hundred to a couple of thousand bases upstream of the TSS. Therefore 2 kb upstream of p110δ exon -2b in both mouse and human were aligned. The stringency was set at more than 5 bp and over 90% homology. 14 homologous regions were identified (Figure 7-3). Six transcription factor binding sites were identified (Table 6-2) which we regarded as possible regulatory factors involved in p110δ expression.

To try and classify whether any of these transcription factor binding sites function in the regulation of p110δ protein expression and to try to identify any other possible regulatory elements within this 2 kb region, PCR primers were designed to amplify the area from mouse genomic DNA (Figure 7-1).

Figure 7-1 Diagrammatic representation of the mouse 5'UTR of the p110δ gene. The positions of the PCR primers used to create the initial 2 kb fragment are marked in red.

The PCR product was cloned into the PGL3 luciferase reporter vector (Promega) and the sequence verified using the RV3 and GL2R vector sequencing primers (Figure 7-2).

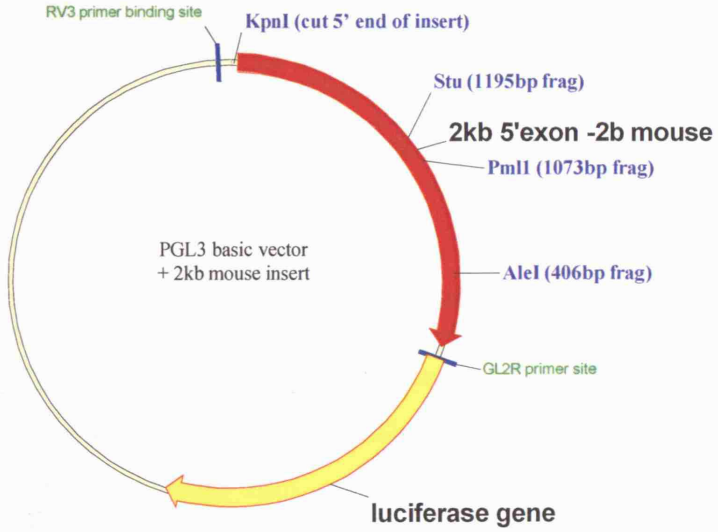

Figure 7-2 Diagrammatic representation of the 2 kb fragment of mouse p110δ genomic DNA cloned upstream of the firefly luciferase reporter gene. The position of the forward and reverse sequencing primers are marked in green. The sites at which the restriction enzymes, used to make deletion constructs, cut are marked in blue.

In an attempt to identify regions with specific promoter activity, smaller fragments were made from the 2 kb fragment using restriction enzyme digests which removed sections of DNA from the 5' end of the insert (Figure 7-2). The lack of restriction enzyme cutting sites limited the number of deletion fragments that could be made easily using this method. Therefore another set of PCR primers were designed to amplify a 150 bp fragment. A 58 bp oligonucleotide was designed to represent the 58 bp of the intron sequence immediately 5' of exon -2b (Figure 7-3). The smallest two fragments (150 bp and 58 bp) did not contain any of the homologous regions.

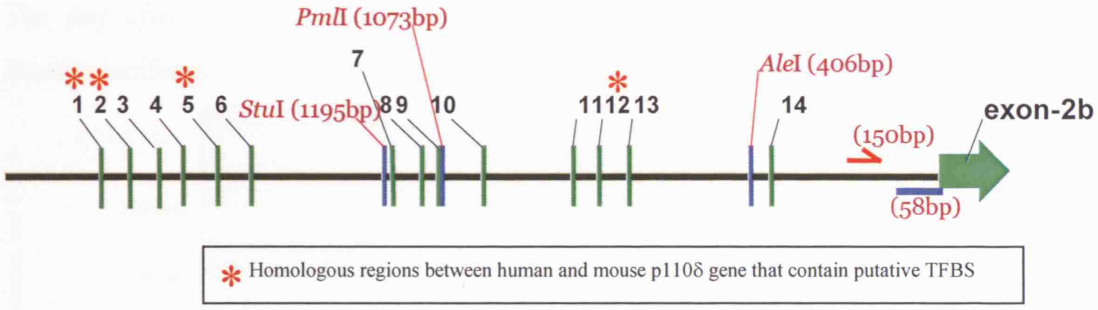

Figure 7-3 Diagrammatic representation of the 2 kb fragment of mouse genomic DNA upstream of exon -2b. The 14 regions of homology between human and mouse DNA recognized using LAGAN pairwise alignment software (Table 6-2) are numbered 1-14. The red arrow represents the position of the PCR primer designed to amplify the 150 bp region of the genomic DNA immediately upstream of mouse p110δ exon -2b. The reverse primer used was complementary to an internal region of the exon. The blue horizontal line represents the position of the oligonucleotide designed to mimic the 58 bp region of genomic DNA immediately upstream of exon -2b used in this set of experiments as the smallest deletion construct. The numbers in brackets are the size of the fragments produced after restriction enzyme digestion or from the PCR amplification. The red stars highlight the areas of homology between the mouse and human DNA that contain putative transcription factor binding sites.

7.2.1.1 Transient transfection of luciferase reporter constructs containing potential tissue-specific elements upstream of exon -2b of the p110δ gene

$3x10^5$ NIH3T3 cells and $3x10^6$ A20 cells were seeded per 60 mm dish the day before transfection. NIH3T3 cells were transiently transfected, in triplicate, using the Qiagen Superfect method and the A20 cells were electroporated. The Promega Dual Luciferase reporter assay was used to allow the difference in transfection efficiencies between two cell lines to be taken into account. One plate of cells from each cell line was transfected with an empty vector (PGL3 basic) to enable the results to be normalized.

As a positive control we used reporter constructs for mouse p110γ. The first one contained 1232 bp of DNA 5' of the transcription start site, previously demonstrated to have leukocyte-specific promoter activity (174). The second one contained 557 bp of DNA 5' of the transcription start site, previously demonstrated not to have any promoter activity (174). The p110γ regions of interest (1232 bp and 557 bp) were cloned in front of the firefly luciferase gene in the PGL3 vector so that they could be used as direct and comparable controls in our transfection experiments. All six p110δ gene constructs, the positive p110γ controls and the empty vector control were co-transfected with the *Renilla* luciferase gene, under the control of the SV40 promoter.

The day after transfection the cell samples were lysed and assayed for firefly and *Renilla* luciferase activity. The counts were normalized and compared (Figure 7-4).

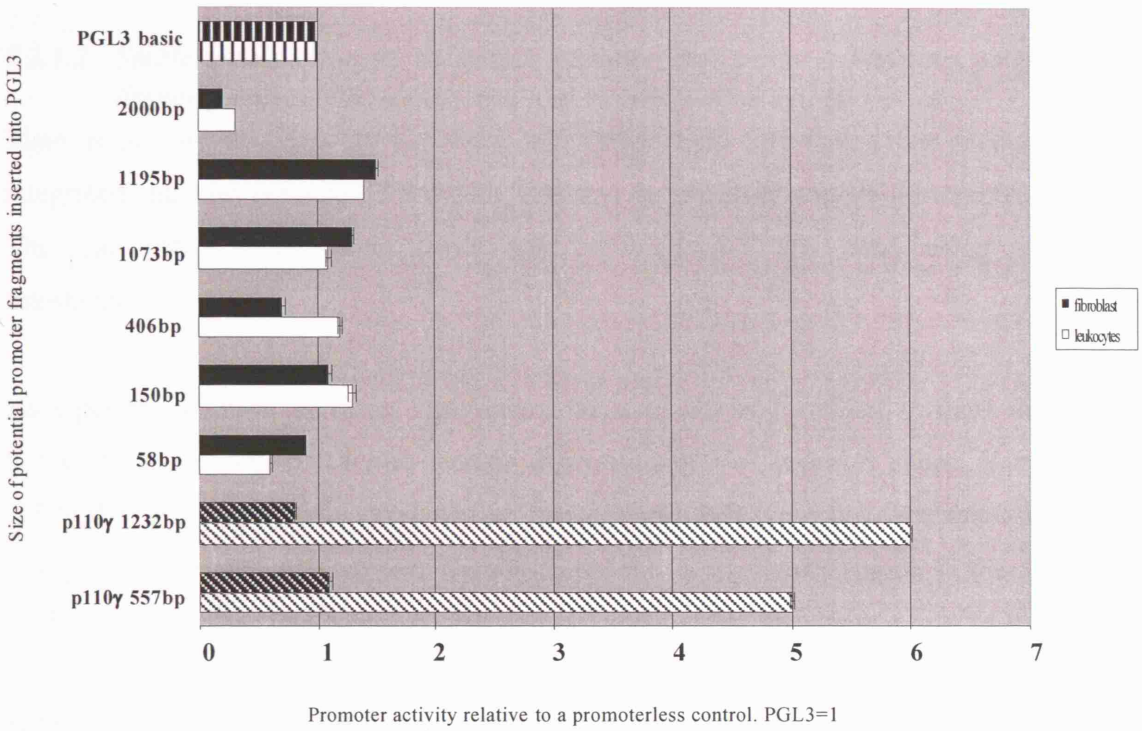

Figure 7-4 Promoter activity of each of the potential promoter elements in two different cell lines as determined by the Promega Dual luciferase reporter assay. The size of each construct generated to study the promoter activity of the murine p110δ gene and their relative luciferase activity is shown. The vertical striped bars represent the activity of the promotorless control. The diagonal striped bars represent the luciferase activity of the p110γ gene constructs. n = 3

Transcript 3 (incorporating exon -2b) has not been identified in fibroblasts to date. It has however, been identified in leukocytes therefore we expected to see some, if not all, of the constructs demonstrating promoter activity in the A20 cells but not in the NIH 3T3 cells. No significant promoter activity was produced from any of the p110δ experimental reporter constructs in either the leukocyte or non-leukocyte cell line.

The 1232 bp positive control confirmed tissue-specific expression of the p110γ gene construct in a mouse leukocyte cell line. However, the 557 bp p110γ control, previously reported to have no promoter activity (174), demonstrated significant tissue specific promoter activity in our hands. Hirsch *et. al.* (174) transfected their mouse p110γ constructs into human leukocyte (U937) and non-leukocyte (Hela) cell lines. The regulatory mechanism involved in the reduction of reporter activity from

the 557 bp construct in human cells may not be active in the mouse cells used in our experiments.

7.2.1.2 Stable transfection of luciferase reporter constructs containing potential tissue-specific elements upstream of exon -2b of the p110δ gene

Many regulatory elements do not work when transiently transfected but need to be integrated into the genome of the cell. This can be achieved via stable transfection. The transfection experiment above was repeated but this time using stable transfection.

As a positive control for these experiments we used a Vav promoter element that is known to drive leukocyte tissue-specific expression of Vav reporter genes. Vav is a GTP/GDP exchange factor expressed in haemopoietic cell types but very few others. Using a reporter assay it has been demonstrated that a regulatory region in the 5'end of the Vav gene has leukocyte-specific activity when stably transfected into EL4 cells (mouse T cell leukaemia). The same region of the Vav gene stably transfected into NIH 3T3 cells (mouse fibroblasts) did not display promoter activity (124). This reporter construct was obtained from Dr Dimitris Kioussis (MRC National Institute for Medical Research, Mill Hill). We cloned the region of interest in front of the firefly luciferase gene in the PGL3 reporter vector.

The A20 and NIH3T3 cells were co-transfected with the firefly experimental vector and a selection vector containing a neomycin resistance gene. The addition of neomycin allowed for the selection of successfully transfected cells.

All of the stably transfected NIH3T3 cell samples grew successfully in media plus neomycin. In contrast only three out of the six experimental constructs were successfully stably transfected into the A20 cells. Therefore only the latter three constructs have been compared between cell types (Figure 7-5). Determination of the tissue-specificity of the vav regulatory region was performed by measurement of luciferase activity under the control of the SV40 viral promoter for normalisation. Ogilvy *et al.* produced stable transfectants by coelectroporation of a mouse hematopoietic cell line (FDC-P1) or lipofection of a mouse fibroblast cell line (NIH3T3) with the vav-β-gal plasmid and the pSV2neo plasmid (incorporating the

neomycin resistance gene) (124). We replicated their method using our cell lines, A20 and NIH3T3 (Figure 7-5A).

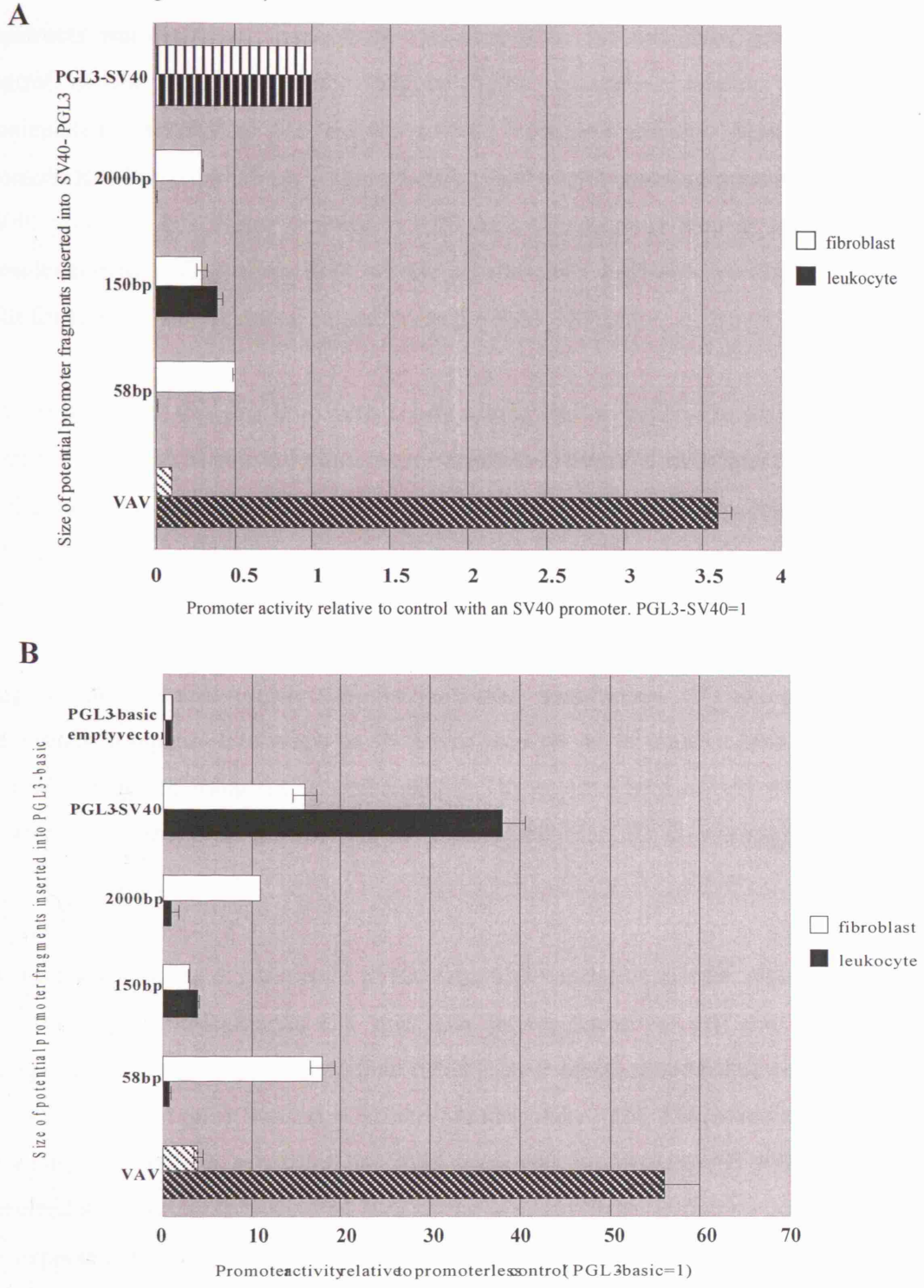

Figure 7-5 Analysis of promoter activity of 3 potential promoter elements upstream of exon -2b of the p110δ gene. (A) The promoter activity of the three potential promoter elements and the PGL3 vector under the control of the SV40 viral promoter was normalised to the luciferase activity produced from the PGL3 promoterless vector. **(B)** The promoter activity of the three potential promoter elements was normalised to the PGL3 promoterless vector. The leukocyte-specific Vav promoter construct was used as a positive control. Note the very different scales of the two graphs. n = 3

The stable transfections were successful, as assessed by the tissue-specific expression of the Vav reporter construct. The reporter activity produced from all of the p110δ constructs was less than the activity recorded from the luciferase gene under the control of the SV40 promoter (Figure 7-5A). Therefore relative activity was recalculated normalizing the results to the recorded activity from the empty promoterless PGL3 construct (Figure 7-5B). This recalculation demonstrated that the SV40 promoter has higher activity in A20 cells (leukocytes) than in NIH3T3 cells (non-leukocytes). This means that we use an artificially high denominator in the A20 cells for further calculation, as compared to the NIH 3T3 cells.

Two solutions to this problem were (1) to normalize the luciferase activity recorded from the p110δ experimental constructs transfected into A20 cells and NIH3T3 cells to the luciferase activity recorded from the SV40 construct transfected into either the A20 cells or the NIH3T3 cells, or (2) to normalize the experimental results to a promoterless control transfected into the corresponding cell line.

Due to the different transfection methods the transfection efficiencies were not sufficiently comparable to allow us to normalize one set of experimental results to a control transfected using the second method. The experimental results were therefore normalized to the promoterless control transfected into the corresponding cell line (Figure 7-5B).

Two out of the three experimental p110δ reporter constructs showed higher promoter activity in the non-leukocyte cell line than in the leukocyte cell line. This is an unexpected result as experiments done to date have demonstrated that non-leukocyte cell lines do not express transcript 3 (incorporating exon -2b). This could be due to the position in which the construct has integrated into the genome. If integration has occurred at a less tightly regulated area of the genome higher promoter activity would be expected. However if this were the case the same scenario would apply to the regulation of the construct in the A20 cells.

A drawback to stable transfections is the lack of knowledge as to where in the genome the transfected genes have integrated. A result showing little or no promoter activity

might suggest integration has occurred in a tightly regulated region of the genome. The higher activity of most of the constructs in the fibroblast cell line could also suggest the absence of a repressor element in the constructs, which is normally present *in vivo*. Alternatively the constructs may have integrated into a region of the genome that is actively transcribed, in contrast to the activity of the promoterless control. There is also the possibility that *in vivo* the chromatin structure surrounding exon -2b in low p110δ expressing cells, such as the fibroblast, is tightly packed therefore preventing the transcription machinery accessing the DNA. This may not be the case in stably transfected cells.

The scenario surrounding stable transfection applies to all of the stably transfected constructs, including the Vav positive control; therefore it is difficult to make a firm conclusion about the promoter activity reported here as there is very little difference in the p110δ promoter activity between the two cell lines tested in comparison to the significant difference seen with the Vav positive control.

7.2.2 Investigation of potential regulatory elements upstream of exon -2a, exon -1 and exon 1 in the p110δ gene

The only tissue-specific difference identified, so far, is the presence of an extra transcript, containing exon -2b, in cells of leukocyte origin. However, two other transcripts were also identified from the 5'RACE experiments (Chapter 5). Evidence of these transcripts found in databases confirms these findings (Table 5-2).

All three p110δ mRNA transcripts have a different first exon. Evidence to date suggests that alternative splicing only occurs in internal exons of mRNAs and that first exons of an mRNA are not spliced out. This suggests that the three-p110δ transcripts are transcribed from three different promoters.

Extensive work to try and identify a positive regulatory region upstream of exon -2b was unsuccessful. In retrospect too much work was performed on the "non-exon -2b region" of the p110δ gene in an attempt to identify regulatory regions upstream of exon -2a, exon -1 and exon 1 in the 5'UTR of the mouse p110δ gene.

7.2.2.1 *Assessment of potential regulatory regions upstream of exon -2a of the p110δ gene*

A 2 kb region upstream of exon -2a was amplified from mouse genomic DNA by PCR (Figure 7-6) and cloned into the PGL3 luciferase reporter vector. A LAGAN mouse/human homology search identified 11 regions of homology of over 5 bp with over 90% homology (Figure 7-7). Within three of the regions of homology transcription factor binding sites were identified as possible targets for the regulation of p110δ expression (Table 6-3).

Figure 7-6 Diagrammatic representation of the mouse 5'UTR of the p110δ gene. The positions of the PCR primers used to create the initial 2 kb fragment are marked in red.

Restriction enzyme digestion and PCR amplification were used to create deletion constructs to try and identify possible regulatory elements and to determine whether any of the homologous regions are required for transcription activation in mouse leukocyte and fibroblast cell lines. In total 7 constructs were created with putative promoter inserts ranging from 2 kb to 36 bp in size: 2 kb, 1500 bp, 970 bp, 742 bp, 479 bp, 256 bp and 36 bp (Figure 7-8).

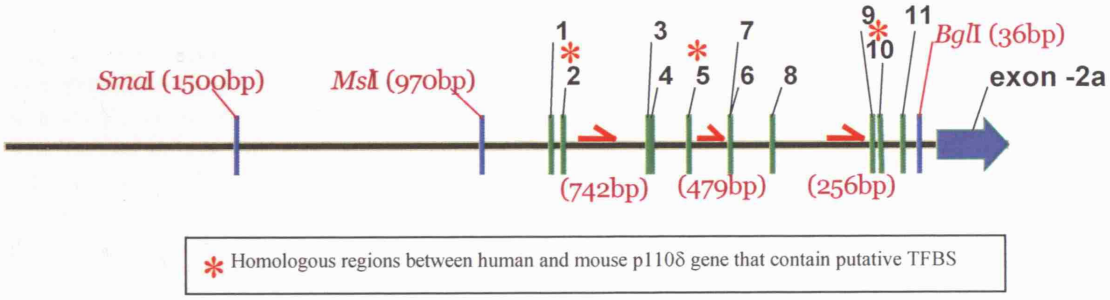

Figure 7-7 Diagrammatic representation of the 2kb fragment of mouse genomic DNA upstream of exon -2a. The 11 regions of homology between human and mouse DNA recognized using LAGAN pairwise alignment software (Table 6-3) are numbered 1-11. The red arrows represent the position of the PCR primers designed to amplify smaller regions of the genomic DNA. The reverse primer used was complementary to an internal region of the exon. The numbers in brackets are the size of the deletion fragment produced after restriction enzyme digestion or from the PCR amplification. The red stars highlight the areas of homology between mouse and human DNA that contain putative transcription factor binding sites.

7.2.2.1.1 Transient transfection of luciferase reporter constructs containing potential regulatory elements upstream of exon -2a of the p110δ gene

NIH3T3 and A20 cells were transiently co-transfected, in triplicate, with the experimental reporter vectors shown in Figure 7-7 and the *Renilla* luciferase vector and assayed for luciferase activity (Figure 7-8).

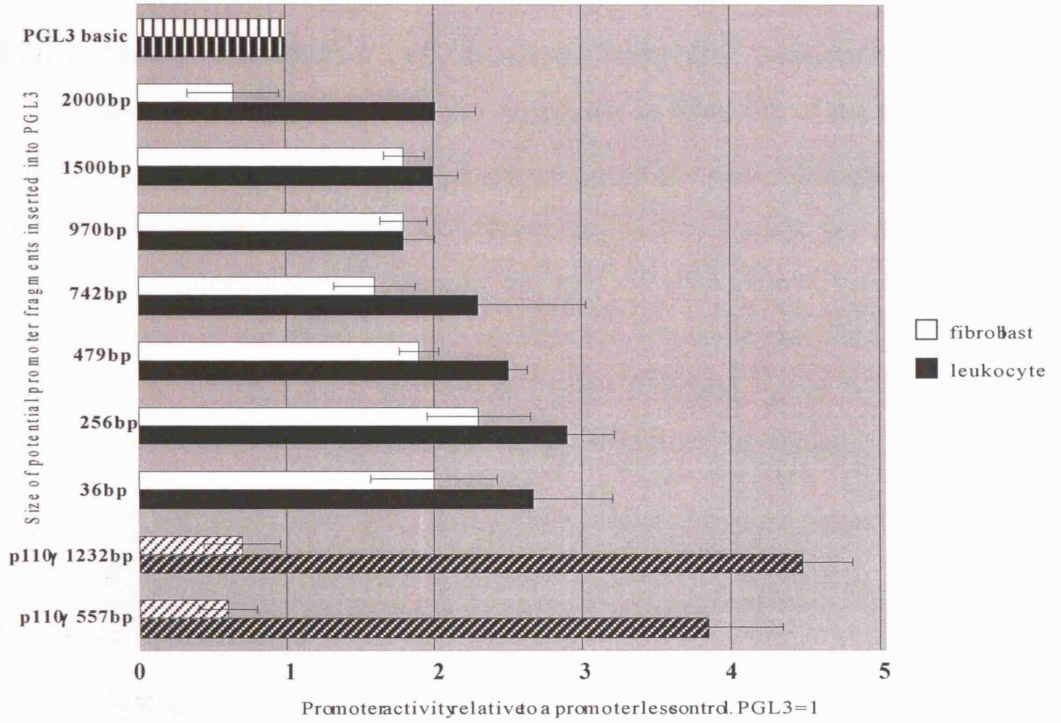

Figure 7-8 Promoter activity of each of the potential promoter elements in two different cell lines as determined by the Promega Dual luciferase reporter assay. The size of each construct generated to study the promoter activity of the murine p110δ gene and their relative luciferase activity is shown. The vertical striped bars represent the activity of the promotorless control. The diagonal striped bars represent the luciferase activity of the p110γ gene constructs. n = 3

The 1232 bp positive control confirmed tissue-specific expression of the p110γ gene construct in a mouse leukocyte cell line. However, we still recorded significant tissue specific promoter activity from the 557 bp p110γ control, previously reported to have no promoter activity (174). A two-fold increase in reporter activity was recorded from most of the constructs, compared to the promoterless control (Figure 7-8). No region within the 2 kb fragment upstream of exon -2a could be identified as a potential regulatory element as all of the constructs demonstrated some promoter activity, even the 36 bp fragment, in both the leukocyte and fibroblast cell lines.

There is no *in silico* evidence, such as regions of homology or binding sites for regulatory elements, to suggest that the 36 bp region immediately upstream of exon -2a is important for the transcription of the -2a transcript. Without doing further deletion studies it is not possible to determine whether this 36 bp region contains a regulatory element for the control of p110δ gene expression.

7.2.2.1.2 Stable transfection of luciferase reporter constructs containing potential regulatory elements upstream of exon -2a of the p110δ gene

All of the seven constructs containing fragments from the genomic region upstream of exon -2a were successfully stably transfected into NIH3T3 cells, whereas only three of these were successfully stably transfected into A20 cells. The obvious difference in promoter activity of the SV40 reporter construct in leukocyte and fibroblast cell lines, as mentioned in section 7.2.1.2 and seen in Figure 7-5B and Figure 7-9 below, makes it an unsuitable control with which to normalize the rest of the results.

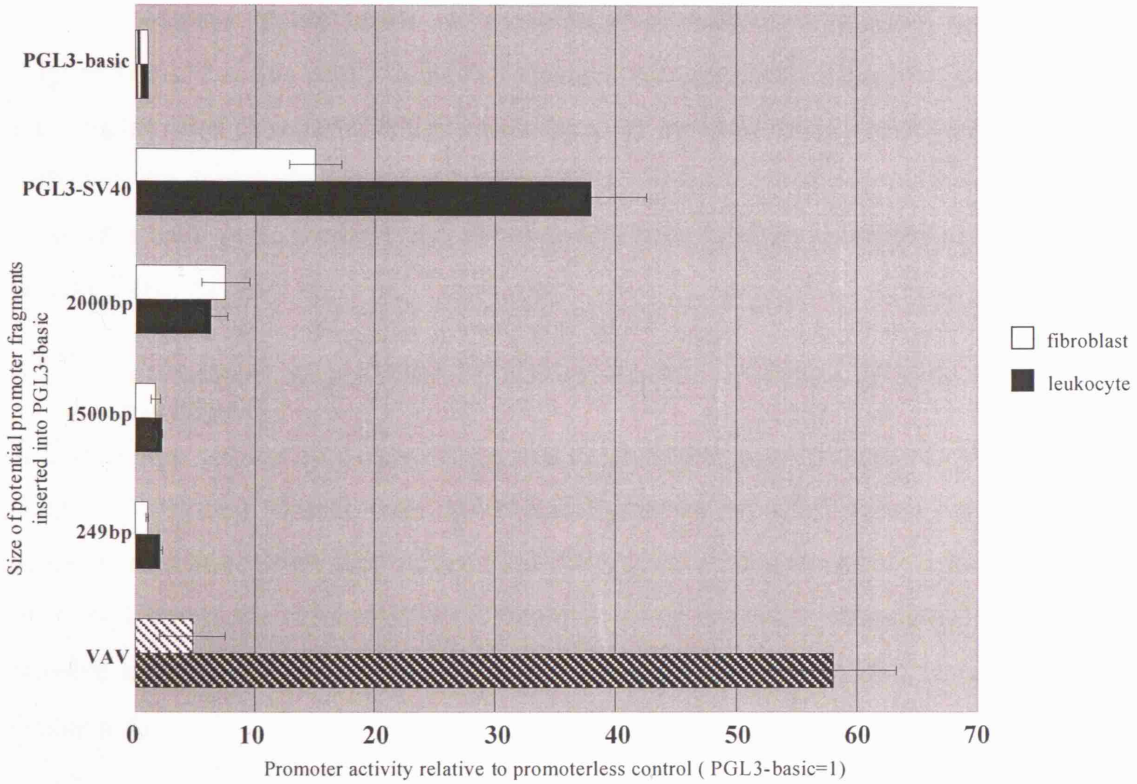

Figure 7-9 Analysis of promoter activity of 3 potential promoter elements in upstream of the p110δ exon -2a normalised to the PGL3 promoterless reporter vector. The leukocyte-specific Vav promoter construct was used as a positive control. n = 3

Therefore the luciferase activity recorded from the stably transfected constructs, containing DNA fragments from the genomic region upstream of exon -2a of the p110δ gene, were normalized to background luciferase activity recorded from the PGL3 promoterless control construct and not to the luciferase activity recorded from the SV40 reporter construct (Figure 7-9).

Evidence of low Vav promoter activity in fibroblast cell lines suggests integration occurred into a less tightly controlled area of the genome than in *in vivo* situations. A significant drop in promoter activity between the 2000 bp fragment and the 1500bp fragment, in both the fibroblast and leukocyte cell lines, suggests this 500 bp region may contain a regulatory element. This was not evident in the *in silico* analysis. More detailed analysis of this 500 bp region is required before further conclusions can be drawn.

Real time PCR analysis using transcript-specific primers demonstrated that leukocyte cell lines express higher levels of transcript 2 as well as expressing transcript 3 (Figure 5-20). For this result to be corroborated using reporter assays we expected to see a higher level of reporter activity from some of the constructs transfected into A20 cells than the constructs transfected into NIH3T3 cells. This was not evident from the reporter assays performed on the putative regulatory region upstream of exon -2a (Figure 7-9).

7.2.2.2 Assessment of potential regulatory regions upstream of exon -1 of the p110δ gene

A 2 kb region upstream of exon -1 was amplified from mouse genomic DNA by PCR (Figure 7-10) and cloned into the PGL3 luciferase reporter vector. A LAGAN mouse/human homology search identified 37 regions of homology of over 8bp with over 90% homology (Figure 7-11). Twelve of these regions of homology contained putative transcription fact binding sites for the regulation of p110δ gene expression (Table 6-4).

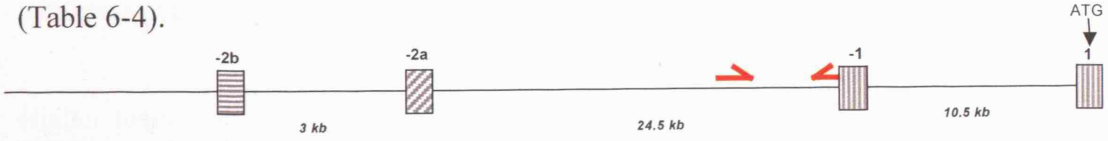

Figure 7-10 Diagrammatic representation of the mouse 5'UTR of the p110δ gene. The positions of the PCR primers used to create the initial 2 kb fragment are marked in red.

Restriction enzyme digestion and PCR amplification were used to create deletion constructs to try and identify possible regulatory elements and to determine whether any of the homologous regions are required for transcription activation in mouse leukocyte and fibroblast cell lines. In total 7 constructs were created with putative promoter inserts ranging from 2 kb to 185 bp in size: 2 kb, 1833 bp, 1685 bp, 1169 bp, 989 bp, 539 bp and 185 bp.

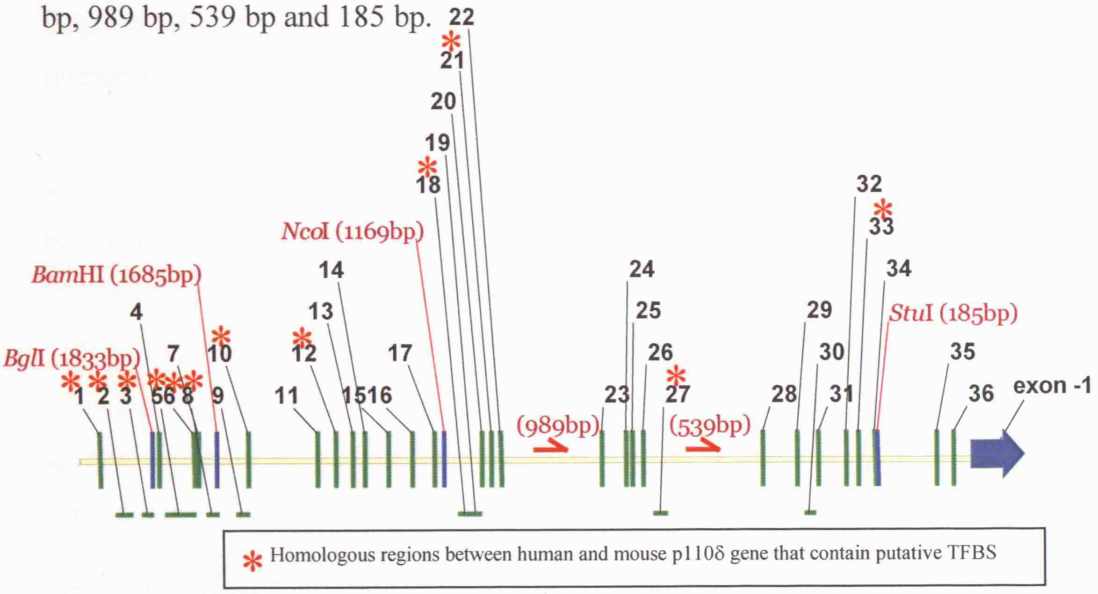

Figure 7-11 A diagrammatic representation of the 2kb fragment of mouse genomic DNA upstream of exon -1. The 36 regions of homology between the human and mouse DNA recognized using LAGAN pairwise alignment (Table 6-4) are numbered 1-36. The horizontal lines represent the homologous sequences over 20 bp in length. The red arrows represent the position of the PCR primers designed to amplify smaller regions of the genomic DNA. The reverse primer used was complementary to an internal region of the exon. The numbers in brackets are the size of the deletion fragment produced after restriction enzyme digestion or from the PCR amplification. The red stars highlight the areas of homology between mouse and human DNA that contain putative transcription factor binding sites.

7.2.2.2.1 Transient transfection of luciferase reporter constructs containing potential regulatory elements upstream of exon -1 of the p110δ gene.

NIH3T3 and A20 cells were transiently co-transfected, in triplicate, with the experimental reporter vectors and the *Renilla* luciferase vector and assayed for luciferase activity (Figure 7-12).

Higher luciferase activity was detected from the genomic region upstream of exon -1 than the genomic region upstream of either of the other two exons, -2a and -2b. There

was over a 3-fold increase in promoter activity from nearly all of the constructs. Two significant results were obtained from this set of experiments;

1) The largest putative promoter constructs were significantly more active in fibroblasts than in B-cells. The reporter activity in B-cells increased to similar levels as those seen in fibroblasts upon the deletion of 315 bp from the 5' end of the p110δ fragment (Figure 7-12). One possible explanation for this is that a tissue-specific repressor element was deleted among the 315 bp of DNA removed.

2) A 12-fold increase in reporter activity, compared to the promoterless control, was recorded in the A20 cell line in contrast to a 3-fold increase in the NIH3T3 cells from the 185 bp p110δ construct (Figure 7-12). The increase in reporter activity from the 185 bp fragment compared to the 539 bp fragment suggests that a region involved in silencing was removed during the deletion process. Therefore there may be a tissue-specific regulatory region within this 354 bp region. Further analysis would have to be carried out to confirm this hypothesis.

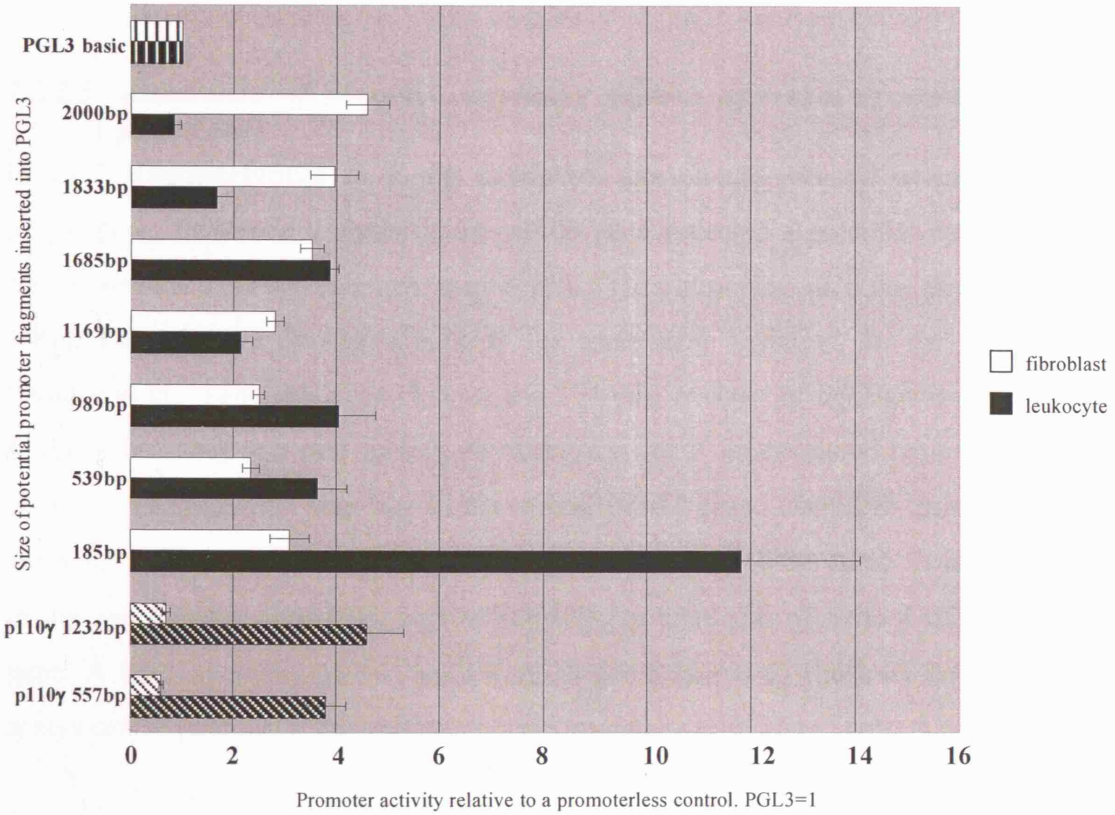

Figure 7-12 Promoter activity of each of the potential promoter elements in two different cell lines as determined by the Promega Dual luciferase reporter assay. The size of each construct generated to study the promoter activity of the murine p110δ gene and their relative luciferase activity is shown. The vertical striped bars represent the activity of the promotorless control. The diagonal striped bars represent the luciferase activity of the tissue-specific p110γ gene constructs. n = 3

No tissue-specific differences upstream of exon -1, between leukocyte and non-leukocyte cell lines, have been identified previously. A number of breast cell lines were identified as expressing transcript 1 (incorporating exon -1 alone) (Table 5-3). No breast cell lines were used in the reporter assays therefore the tissue-specific reporter activity demonstrated above from certain p110δ constructs was unexpected.

7.2.2.2.2 Stable transfection of luciferase reporter constructs containing potential regulatory elements upstream of exon -1 of the p110δ gene

Only the 1685 bp, 1169 bp and the 989 bp constructs were successfully transfected into the A20 cell lines. The promoter activity was low compared to the PGL3 SV40 construct (data not shown). Without the 2 kb construct, 539 bp construct and 185 bp construct fragment stably transfected into A20 cells we were unable to determine whether the significant difference in reporter activity between the two set of constructs, described above, is reproducible in stable transfections.

7.2.2.3 Assessment of potential regulatory regions upstream of exon 1 of the p110δ gene

In 1997 Vanhaesebroeck *et al.* (8) published data on the primary structure of the p110δ gene. Sequence analysis of the p110δ gene revealed a potential open reading frame, preceded by an in-frame stop codon. They reported that the putative start codon lies within a favorable context for translation initiation. In the same year Chantry *et al.* (188) published data on the 5'RACE analysis of the mouse and human p110δ gene. They reported finding 36 nucleotides of 5' untranslated region upstream of the ATG translation start site in the mouse p110δ gene. Therefore, prior to doing the 5'RACE experiments, thus prior to the identification of the three 5' untranslated exons, we focused on a 2 kb region immediately upstream of exon 1 of the p110δ gene. A brief overview of the work done is given below to illustrate how reporter assays can be potentially misleading.

A 2 kb region, upstream of exon 1, was amplified from mouse genomic DNA by PCR (Figure 7-13) and cloned into the PGL3 luciferase reporter vector. A LAGAN mouse/human homology search identified 10 regions of homology of over 8 bp with

over 90% homology (Figure 7-14). Three putative transcription factor binding sites possibly involved in the regulation of p110δ gene expression were identified within three of these homologous regions (Table 6-5)

Figure 7-13 Diagrammatic representation of the mouse 5'UTR of the p110δ gene. The positions of the PCR primers used to create the initial 2 kb fragment are marked in red.

Restriction enzyme digestion and PCR amplification were used to create deletion constructs to try and identify possible regulatory elements and to determine whether any of the homologous regions are required for transcription activation in mouse leukocyte and fibroblast cell lines. In total 9 constructs were created with putative promoter inserts ranging from 2 kb to 56 bp in size: 2 kb, 569 bp, 446 bp, 254 bp, 185 bp, 117 bp, 91 bp, 64 bp and 56 bp (Figure 7-14).

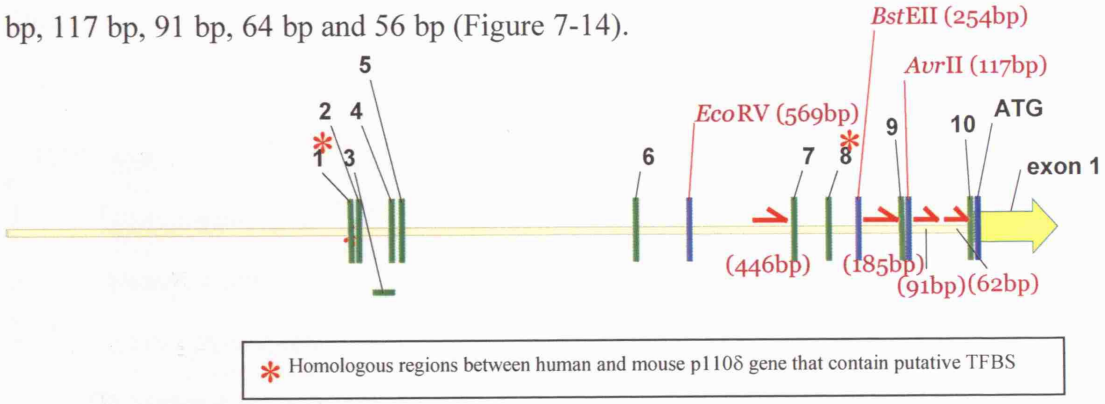

Figure 7-14 A diagrammatic representation of the 2 kb fragment of mouse genomic DNA upstream of exon 1. The 10 regions of homology between the human and mouse DNA recognized using LAGAN pairwise alignment are numbered 1-10. The horizontal line represents the homologous sequence over 20 bp in length. The red arrows represent the position of the PCR primers designed to amplify smaller regions of the genomic DNA. The reverse primer used was complementary to an internal region of the exon. The numbers in brackets are the size of the deletion fragment produced after restriction enzyme digestion or from the PCR amplification. The red stars highlight the areas of homology between mouse and human DNA that contain putative transcription factor binding sites (Table 6-5).

7.2.2.3.1 Transient transfection of luciferase reporter constructs containing potential regulatory elements upstream of exon 1 of the p110δ gene

NIH3T3 cells and A20 cells were transiently co-transfected in triplicate and assayed for luciferase expression (Figure 7-15).

None of the transcripts, identified by 5'RACE, begin at exon 1. Although this suggests that there is not a promoter in this region it does not rule out the possibility of the presence of regulatory elements.

The reporter activity recorded in both the leukocyte and fibroblast cells was significantly higher than previously recorded activity. A region of 192 bp (-64 to -254) recorded a greater than 20-fold increase in reporter activity compared to the promoterless control. This was the case in both the fibroblast and leukocyte cells. The activity decreased to just over 2-fold when a further 6 bp were deleted from the fragment. The lack of *in silico* evidence of a putative promoter in this region suggests that it may contain an enhancer. Further analysis of this 6 bp sequence failed to identify a binding site for a transcription factor. One possible hypothesis is that this small region, in partnership with a region further downstream may be required for the 20-fold increase in reporter activity.

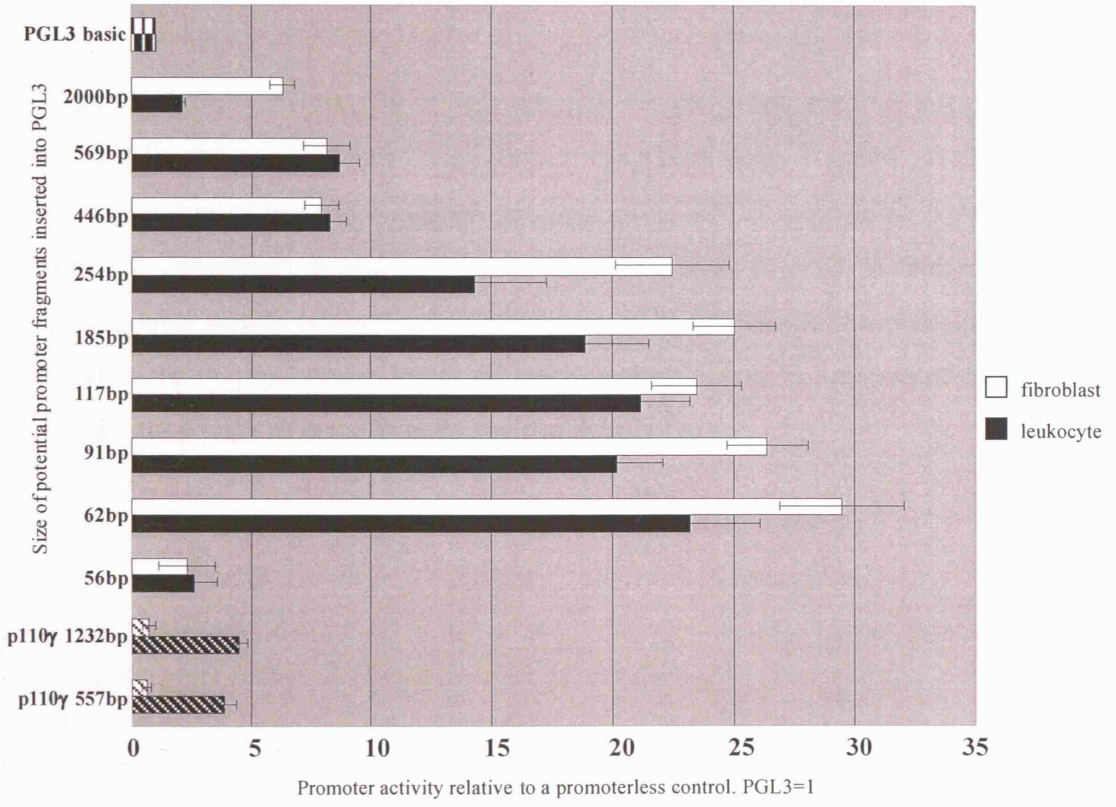

Figure 7-15 Promoter activity of each of the potential promoter elements in two different cell lines as determined by the Promega Dual luciferase reporter assay. The size of each construct generated to study the promoter activity of the murine *PI3K*δ gene and their relative luciferase activity is shown. The vertical striped bars represent the activity of the promotorless control. The diagonal striped bars represent the luciferase activity of the tissue-specific p110γ gene constructs. n = 3

A construct with a further 13 bp deleted yielded no reporter activity suggesting that the 20 bp, -44 to -54, from the ATG translation start site are in some way involved in the regulation of p110δ gene expression. A number of transcription factor binding sites were identified in the 62 bp region upstream of the translation start site of exon 1. Constructs created with mutations in these sites failed to identify any particular elements that may be involved in the regulation of the p110δ protein expression (data not shown).

A select sample of constructs, from DNA regions upstream of each of the four p110δ exons studies, were transiently transfected into another mouse leukocyte cell line, WEHI 231, and two different mouse epithelial cell lines, CT26 and LLC. The results obtained from the WEHI 231 cell line reproduced the results achieved in the A20 cell line. Similarly the CT2 and LLC results followed suit, reproducing the data obtained from the NIH3T3 cells (data not shown). This confirms that the data is reproducible between cell lines.

Without further analysis, it is not possible to pin-point specific sites which can account for the tissue-specific regulation of the p110δ gene. It is not yet clear whether the expression of transcript 3 (incorporating exon -2b) is responsible for the increase in p110δ expression in leukocytes, or whether it is as a result of leukocyte-specific transcription factors acting further downstream, or whether it is a combination of the two factors. A third possibility is silencing factors acting in non-leukocyte cell lines. At this stage none of these hypotheses can be ruled out.

8 Discussion

8.1 Investigation of a possible correlation between p110δ mRNA and protein levels

Leukocytes consistently express high levels of the p110δ protein (7,8,188). Some cell lines of non-leukocyte origin, such as melanocyte/melanoma, small cell lung carcinoma, and normal and breast cancer cells can also express p110δ protein but they tend to express this protein a lower and more variable level than leukocytes (7,8,214). Using improved methods of protein detection we have now been able to detect p110δ in all cell types but often at a very low level (Figure 3-3).

In an investigation into the level at which p110δ is regulated we identified that in mouse cells, there appears to be a rather good correlation between the levels of p110δ mRNA and protein (Figure 3-6), indicating that the regulation of p110δ protein expression mainly occurs at the level of transcription or chromatin regulation and not at a later stage such as translation or protein degradation. In human cells, this correlation did not seem to be as strict. Why human would be different to the mouse in this respect is unclear, and more work needs to be done to be able to draw firm conclusions, both in the mouse and human. Indeed, we may not have investigated enough cells with intermediate levels of p110δ protein expression (Fig 3-6). In future experiments, it will be critical to still better standardize cell culture and harvest, as well as RNA isolation and storage, as we have found that slight changes in these conditions can have a big impact on the results of quantitative mRNA and protein analysis.

The studies described above were carried out in established cell lines, in which the regulation of the p110δ gene may have been altered by the transformation that these cells have undergone. We therefore next tested whether there was a correlation between p110δ mRNA and p110δ protein levels in primary mouse organs. Perfusion of the mice to flush out the circulating leukocytes was carried out, as this apparently could largely overcome the problem of leukocyte contamination of tissues in other studies. As expected, high levels of p110δ mRNA and protein were identified in both

the spleen and thymus, given that these organs mainly consist of leukocytes. Intermediate levels of p110δ mRNA and protein were identified in the lung. Low levels of p110δ protein and mRNA were also identified in the liver, kidney, heart, brain and cerebellum. In some of these tissues, this was most likely due to the presence of leukocytes, as demonstrated by the expression of CD45 (a leukocyte marker) in lysates of liver, lung and kidney samples. However, in other tissues such as brain, cerebellum and heart, in which p110δ protein and mRNA could be detected, a no CD45 signal was seen.

Most likely, p110δ signals in the CD45-positive tissues such as liver and lungs are due to presence of tissue-resident leukocytes such as macrophages which could not be removed by perfusion of the mice. These macrophages differ in appearance in various tissues and are known by various names: Kupfer cells in the liver, reticular cells in the lymph nodes, spleen and bone marrow and alveolar macrophages in the alveoli of the lungs. It is possible that the very low levels of p110δ mRNA and protein in the heart and kidney are also due to tissue-resident macrophages.

The p110δ mRNA and protein found in the brain and the cerebellum was not due to leukocyte contamination, as assessed by CD45 analysis. Unpublished data from our laboratory shows high levels of p110δ in the brain and spinal column in a transgenic p110δ reporter mouse line (Figure 3-9). This is believed to be due to high levels of p110δ expression in neurons.

By and large, there was a rather good (but not perfect) correlation between p110δ mRNA and p110δ protein in primary tissues from the mouse. Again, more rigorous sample preparation and repeat experiments are required to be able to firmly state that there is an absolute correlation between the levels of p110δ mRNA and protein. Taken together, however, our data suggest that post-transcriptional mechanisms do not play a large part in the regulation of p110δ protein expression.

8.2 Induction of p110δ expression

Recent publications have reported an increase in p110δ protein expression in cardiac fibroblasts upon stimulation with a mineralocorticoid (199). Array data characterizing gene expression in resting blood B-cells also reported an increase in p110δ mRNA expression after stimulation with anti-IgM antibody (198). Therefore an attempt was made to upregulate p110δ gene and protein expression in cell lines, such as NIH3T3 and L929, which do not normally express high levels of the p110δ protein, and also to further increase p110δ levels in cells which already express substantial levels of p110δ (such as B cells).

Cells responding to physical, chemical and biological stresses induce protective proteins. Given the implication of PI3K activity in the regulation of cell survival, we considered that cells might upregulate p110δ protein in order to protect themselves during stress responses. Three stress stimuli (TNF-α, UV radiation and sorbitol) failed to increase p110δ protein expression. However, appropriate positive controls were not included in these experiments, and we are therefore not in a position to draw firm conclusions from these experiments.

We considered that p110δ protein might be absent from some cells due to constitutive degradation of the p110δ protein and therefore aimed to interfere with the cellular machinery for protein degradation. However, no changes in the levels of p110δ protein were observed when the cells were incubated with PS-341, an inhibitor of the cellular proteasome. p27, a cyclin-dependent kinase inhibitor known to be regulated by the ubiquitin-proteasome pathway, was stabilized by PS-341 treatment indicating that our experimental conditions were appropriate to induce protein stabilization.

Aldosterone, which has been implicated in cardiac hypertrophy, has been reported to stimulate induction of p110δ protein in cardiac fibroblasts (199). We have not been able to increase p110δ protein expression in NIH3T3 fibroblasts by aldosterone. It would have been advisable to also test cardiac fibroblasts in our experiments, as it is possible that the effect of aldosterone is cell type-specific.

We also failed to increase p110δ protein expression in anti-IgM-stimulated resting primary B-cells. It is possible that the induction of p110δ mRNA as seen in Ref. (198) on primary human B cells did not lead to the expression of the p110δ protein. In a next experiment, we should therefore also investigate the impact of BCR stimulation on p110δ mRNA and protein in parallel. In hindsight, we should also have included a positive control in this experiment, and for example assessed the induction of the Erk/MAPK pathway in the resting B cells, as it is still possible that the BCR stimulation as carried out by us did not lead to effective signal transduction.

The state of chromatin plays an important role in whether a gene is actively transcribed or not. DNA methylation and histone acetylation are responsible for this level of control. Demethylation of DNA loosens a tight chromatin structure allowing transcription factors to bind to specific binding sites in the DNA. Hyperacetylation also leads to chromatin relaxation. To determine whether expression of the p110δ gene is controlled by the state of the chromatin, cells that normally express very low levels of p110δ protein were treated with 5AC (a demethylating agent) and TSA (an inhibitor of histone deacytelases), alone or in combination, to relax the chromatin structure. At the RNA level a two-fold increase in p110δ RNA levels was observed after treatment with 5AC and TSA. This was not reflected in an equal increase in protein expression. Our previous results suggested a correlation between mouse p110δ protein expression and p110δ RNA levels (Figure 3-6) on the basis of which we also expected some increase in p110δ protein expression in these experiments. However, a 2-fold increase may very well be within the error range of our experiments. It is clear though that chromatin modulation appears not to be a major level of control of p110δ gene expression.

Overall, we can conclude that the p110δ gene expression does not appear to be regulated in an acute manner in response to extracellular stimuli.

8.3 Discovery of three distinct transcriptional start sites in the p110δ gene

Isolation and analysis of the 5' end of a cDNA can provide valuable and essential information. First, the transcription start site of a mRNA relative to the corresponding genomic locus can be mapped. This information is essential for the analysis of the promoter of a gene, part of which usually lies immediately upstream of the transcription unit. Second, determining the transcription start site of a gene facilitates investigation of different transcription start sites and alternative first exon usage. Finally, the first coding exon of an mRNA typically includes both a non-coding region upstream of the initiator codon and the translation start site itself. The value of locating the putative start site for translation is obvious, but without knowledge of the 5' untranslated sequence, it is difficult to be certain that 'an' AUG in the cDNA sequence is 'the' AUG start site. The upstream, untranslated region will typically contain stop codons in-frame with the genuine open reading frame, confirming their non-coding identity and supporting the role of a downstream AUG as the genuine translation start site.

Chantry *et al.,* who published the cloning of mouse and human p110δ cDNAs, used 5'RACE to identify the 5' end of the p110δ gene (188). The length of the untranslated region of the human p110δ gene was not reported (188). It was mentioned in their manuscript that the mouse p110δ sequence contained 36 nucleotides of 5' untranslated region but the sequence was not revealed (the sequences they deposited in the database (U86587, also list the other one U86453) all started at the ATG translation start site(188).

Vanhaesebroeck *et al.* had reported the presence of the 5'UTR sequence in human p110δ (8) and had deposited this sequence in GenBank (accession number Y10055). This sequence contained 196 nucleotides 5' of the ATG start codon. Based on the 5'RACE analysis it became clear that this clone represents transcript 2, containing exon -2a (missing 21 bases at the 5' end), exon -1 and exon 1. It contains an in-frame stop codon preceding the ATG, indicating that the full-length coding cDNA had been isolated. This was not realised until later on in my PhD work as all of the initial work was primarily based on the mouse p110δ gene.

5' RACE experiments on p110δ mRNA

Our 5'RACE experiments performed on mRNA obtained from a large panel of cell lines revealed a much larger 5' untranslated region than previously reported, both in the mouse and human p110δ gene.

Out of the 9 mouse and 5 human cell lines and primary cells analysed by 5'RACE, PCR products with three different sizes were identified, in both species. When the sequences of the PCR products were analysed using BLAT genome alignment software, 3 different p110δ transcripts in both mouse and human p110δ gene were revealed (Figure 8-1).

Transcript 1 contains the 36 nucleotides upstream of the ATG translation start site in exon 1 and an extra 120 bp of DNA (now referred to as exon -1) which aligned to a region 10.5 kb upstream of exon 1 on the mouse genomic DNA. A similar alignment was seen for the human p110δ transcript 1, except that the human exon -1 is 18.9 kb upstream of human exon 1 (Figure 8-2).

p110δ transcript 2 in both mouse and human p110δ contains the same nucleotides as transcript 1, with an additional 142 bp and 72 bp, respectively (now referred to as exon -2a). These sequences aligned to a region 24.5 kb and 39.7 kb upstream of exon -1 in the mouse and human p110δ gene, respectively (Figure 8-2). It was not until we had performed the 5' RACE experiments and had identified the 3 distinct mouse p110δ transcripts, that we aligned the human p110δ sequence, published by Vanhaesebroeck *et al.* (8) in GenBank, to human genomic DNA. It was only then that we realised that this sequence was the human homologue of the mouse p110δ transcript 2 (missing 21 bases at the 5' end).

Transcript 3 in both mouse and human contains the same nucleotides as transcript 1 with an additional 89 bp and 251 bp, respectively (now referred to as exon -2b). These sequences align to a region 27.5 kb and 61.4 kb upstream of exon -1 in the mouse and human p110δ gene, respectively (Figure 8-2).

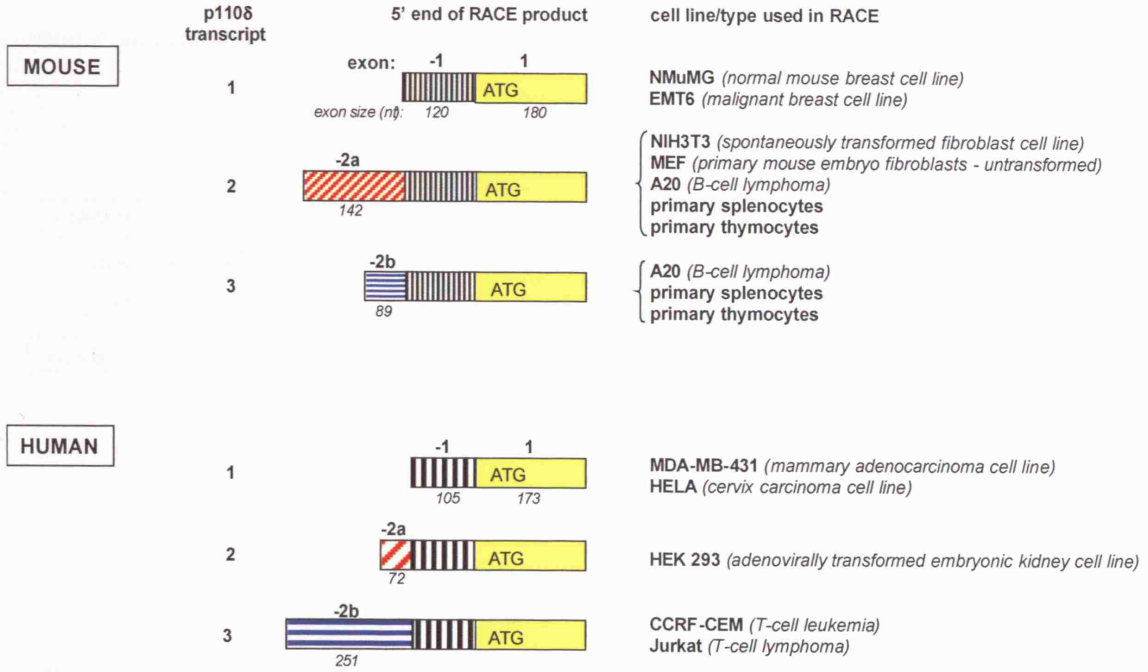

Figure 8-1. Diagrammatic representation of the three distinct 5'UTRs of the p110δ mRNA transcripts in mouse and human, as revealed by 5' RACE

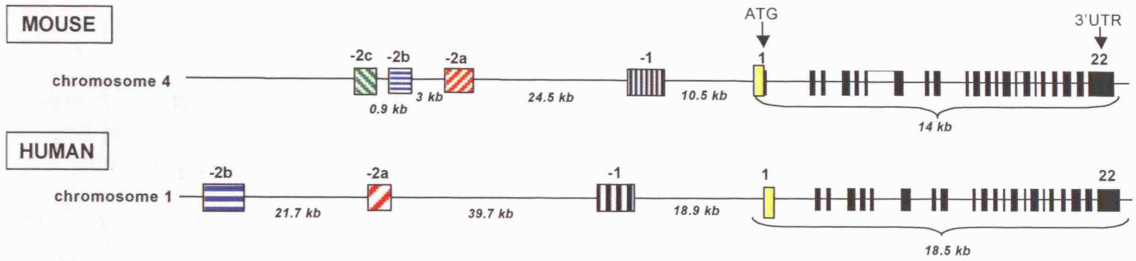

Figure 8-2 Genomic organisation of the mouse and human p110δ genes

Mouse and human exon -1 are highly homologous. No homology was identified between exons -2b and -2a in mouse and human p110δ. Part of this may relate to the considerable size difference between the mouse and human exons (80 *versus* 251 bp (exon -2b) and 142 bp *versus* 72 bp (exon -2a), respectively.

Thus far, no RACE products that incorporate both exons -2b and -2a have been found. This indicates that transcript 3 (incorporating exons -2b/-1) arises from -2b/-2a/-1 transcripts by splicing out of exon -2a. That this is possible is indicated by the presence of a splice donor at the 5' end of the intron between exon -2b and exon -2a

and an acceptor at the 3' end of the intron between exon -2a and exon -1, in both the mouse and human p110δ gene (Figure 8-3).

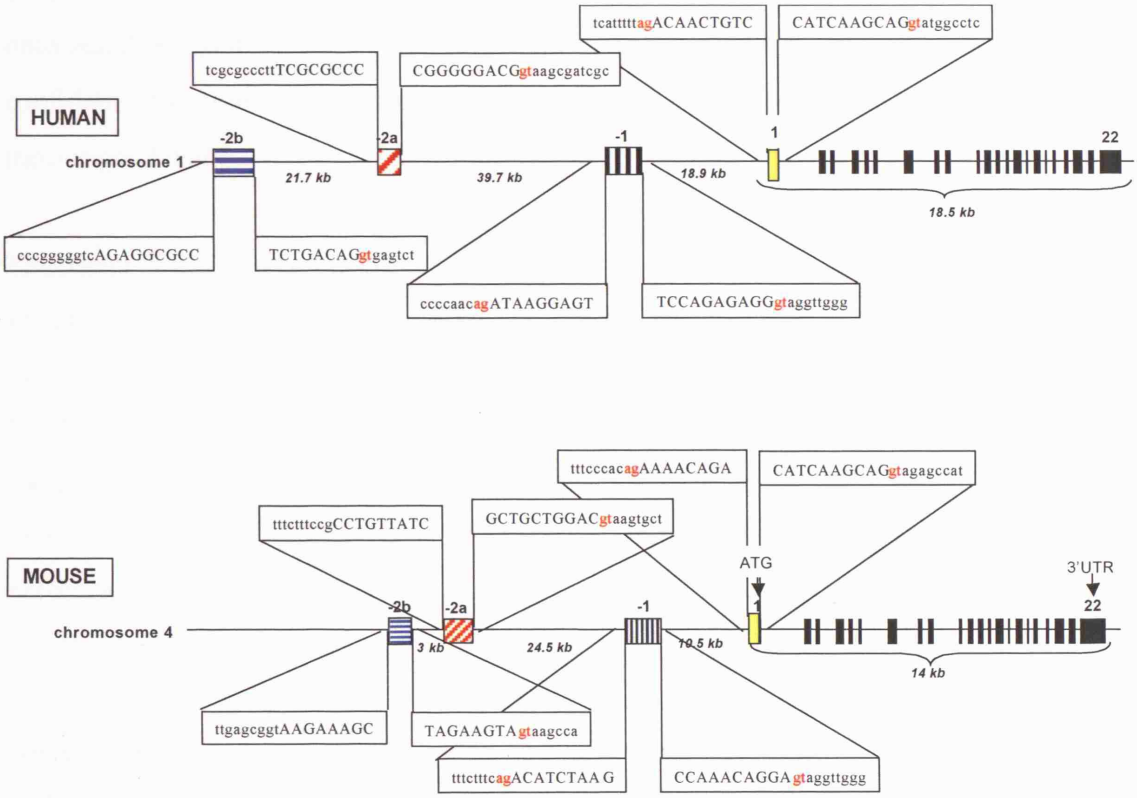

Figure 8-3 Diagrammatic representation of the splice donor (gt) and splice acceptor (ag) sites located in between the human and mouse 5'-untranslated exons of the p110δ gene. Uppercase: exon sequences; lowercase: intron sequences.

The absence of an acceptor at the 3' end of the intron between exon -2b and exon -2a makes it highly unlikely that exon -2a and exon -2b will be re present on the same final mRNA transcript after splicing of precursor mRNA has occurred (Figure 8-3). In other words, when the pre-mRNA of transcript 3 is spliced, exon -2b is joined to exon -1 therefore removing exon -2a from the final mRNA transcript. This helps to confirm the sequences obtained from the 5'RACE experiments, in which exon -2b is joined onto exon -1.

This also raises the question how transcript 2 arises. To the best of our knowledge, the exon that is transcribed first (containing the 5'CAP) is not spliced out from mRNAs. For example, transcript 1, containing exon -1 alone, could not be created from either transcript 2 or transcript 3 having spliced out exon -2a or exon -2b, respectively. This

suggests that transcript 2 arises from an mRNA created from a distinct promoter region upstream of exon -2a. The absence of a splice acceptor site 5' of exon -2b and -2a in the mouse and human p110δ gene suggests that they are unlikely to be spliced onto another exon further upstream of the gene. From this information we can be confident that there is a different promoter that drives the expression of p110δ transcripts 2 and 3.

Transcript 1 has a splice acceptor immediately upstream of the most 5' exon (exon -1). This suggests that this exon could, in principle, be spliced onto 5' upstream exons. In our 5'RACE studies, we have obtained 7 mouse and 5 human copies of transcript 1, which is less frequent compared to the other transcripts we identified (transcript 2: 44 and 10 copies in mouse and human, respectively/ transcript 2: 28 and 12 copies in mouse and human, respectively). This raises the question whether the 5'RACE products with exon-1 we have obtained are partial, and in fact represent mRNA species that had incorporated the upstream exons (exon -2a or -2b) which were no longer represented in the final 5'RACE products. That this could happen is explained below when we look closer at the procedure used for 5'RACE (Figure 2-1).

The first step in 5'RACE analysis involves the treatment of the total RNA from cells with CIP which removes the 5'-phosphate group from all RNA species except those that have a CAP structure. Therefore any degraded RNA transcripts are dephosphorylated and unable to be ligated to the synthetic RNA oligonucleotide adapter. The forward primer used in the PCR reaction is complementary to the adapter molecule. Therefore if the adapter does not ligate to the RNA PCR products will not be made and any partial/degraded mRNA transcripts would not be identified by the assay. After the CIP treatment pyrophosphatase is added to the sample which removes the pyrophosphate bond between the CAP and the rest of the mRNA. This creates a new 5' phosphate onto which the primer is annealed. Degradation of such mRNA can in principle occur at this step between TAP addition and adapter ligation, removing for example the exon sequence that was initially incorporated in the mRNA.

All sequences obtained for transcript 1 did contain the full sequence of exon -1. It would be unlikely that aspecific mRNA degradation would end so specifically at the

boundaries of exon -1. Another argument in favour of the presence of a separate transcript 1 is the fact that we have been unable to demonstrate the presence of exons -2b or -2a in a number of breast cell lines (Table 5-4) by a method independent of RACE, namely PCR using primers specific for exon -2a or -2b. This type of PCR was also carried out on cell lines in which we identified as containing only transcript 1 by 5'RACE (NmuMG and EMT6 mouse cell lines and MDA-MB-431 and HeLa human cell lines).

Database analysis identified 9 mouse and 3 human published sequences that represent transcript 1 of the mouse and human p110δ mRNA, respectively (Figure 5-12). 8 out of the 9 mouse transcripts published were identified using a CAP dependent method indicating that they are most likely full-length sequences. None of the human transcripts were identified using this type method, which suggests that there could be exon sequences further upstream (Table 5-3).

Transcript 1 was identified only in some breast cell lines and in HeLa ovarian carcinoma cells. All the mRNA samples from breast cell lines that we have tested, except the EMT6 and NMuMG, were a gift from Alan Macay, Breakthrough Cancer Research, London. We had no control over the cDNA synthesis, and it is just possible that the cDNA of these breast cell lines may not have been very well extended at the 5'end. It is not clear at the moment if transcript 1 would only be expressed in breast and ovarian cancer cells, as none of the transcript 1 sequences published in the database were from breast cell lines or breast tissue (Table 5-3).

A fourth transcript, containing exon -2c, was identified by 5'RACE in one mouse leukocyte cell line (EL4 thymoma). The position of this putative exon is shown in Figure 8-2. PCR primers designed to amplify this exon from p110δ cDNA did not give rise to a PCR product. However, a positive control was not used to assess the efficiency of the PCR primers we had designed to generate the desired PCR fragment. RNase protection assay also failed to confirm the presence of this exon in p110δ mRNA from EL4 and other cells (see below).

From the initially limited number of cell lines used for 5'RACE analysis, a pattern of distinct 5'RACE products emerged. It appeared that transcripts 2 and 3 were found in

human and mouse leukocyte cell lines and the primary mouse spleen and thymus cells. In contrast, mouse NIH3T3 fibroblasts and human HEK 293 adenocarcinoma non-leukocyte cell lines expressed only transcript 2. Another pattern was found in breast cell lines from human and mouse, as well as HeLa human cervical carcinoma cells, which only expressed transcript 1. Large-scale PCR analysis confirmed this pattern. All leukocyte cell lines tested expressed both transcripts 2 and 3. All non-leukocyte cell lines, with the exception of about half of the breast cell lines analysed, expressed transcript 2 only. The breast cell lines which tested negative for transcripts 2 and 3, expressed only transcript 1.

Published data in a genome database containing information from EST sequences and genome-wide determination of full-length mRNAs provided independent confirmation of the presence of exons -2b, -2a and -1 identified in our study (Figure 5-12 and Table 5-3). Taken together, this indicates that there are (at least) 3 promoters for the p110δ gene.

Leukocytes, which invariably express high levels of the p110δ protein, contain transcript 3 that is not present in other cell lines. Melanoma cell lines, such as B16-BL6, which also express relatively high levels of p110δ protein, apparently achieve this without expressing transcript 3, and instead express the p110δ protein from transcript 2 (which is also present in leukocytes). In other words, non-leukocyte cells which produce intermediate levels of p110δ protein, achieve this not entirely in the same way as leukocytes. It is possible that transcript 2 drives an intermediate level of p110δ protein, whereas for higher-level expression a transcript 3 has to be produced. It is truly remarkable that non-leukocytes never express transcript 3, strongly indicative for leukocyte-specific control of production of transcript 3.

Database analysis of other members of the human and mouse PI3K family identified 5' untranslated exons in all members of the catalytic and regulatory subunits except the PIK3r3 gene (p55γ) (Figure 5-21). The presence of 5' untranslated exons is not an uncommon phenomenon, however all the catalytic subunits of the class IA PI3K genes have untranslated exons over 10 kb upstream of the translation start site. The mouse p110α gene has one untranslated exon over 50 kb upstream of the translation start site (Figure 5-12). No 5' untranslated exons were found in other genes

investigated (Rho/Rac GEF, PDGF, PDGFR, EGFR and IL6). This suggests that multiple transcription start sites long distances from the translation start site is not unique to p110δ but may be a specific element in the regulation of most members of the class I PI3K family.

RNAse protection experiments on p110δ mRNA

We next carried out RNAse protection experiments to verify the configuration of the distinct p110δ mRNAs independent from PCR-based strategies. Three radioactively labelled probes were designed to protect p110δ RNA transcripts containing exon -2c, exon -2b and exon -2a (Figure 5-13). The fact that transcripts 1, 2 and 3 all contain exon -1, made it impossible to assess by RNAse protection whether transcript 1 is a genuine transcript or whether it is a product of degraded transcript 2 or 3. In other words, all of the RNAse protection probes contained probe information for exon -1, which is present in all types of transcripts identified (Fig. 8-1), thus not allowing to assess the presence of transcript 1 specifically. The RNAse protection experiments confirmed that the three leukocyte cell lines tested (A20, WEH 231 and EL4) expressed transcript 2 and 3 (Fig. 5-15). In NIH 3T3 cells, which only expressed transcript 2 as evidenced by 5'RACE and exon-specific PCR, no evidence for the presence of transcript 2 could be found. This is most likely due to the low levels of p110δ mRNA in these cells, not detectable by RNAse protection.

The exon -2c probe designed to protect a 337 bp length of mouse p110δ mRNA instead protected a fragment of approximately 260 bp in WEHI 231, A20 and EL4 leukocytes. This could be explained by the presence of transcript 1, an mRNA transcript containing exon -1 and exon 1, without exon -2c. While this present some evidence for the existence of transcript 1, it is surprising that this transcript was never identified by 5'RACE in the EL4 cells (all 5'RACE products represented transcript 2 or transcript 3). The absence of a fully protected fragment using the -2c probe suggests that exon -2c may have been a cloning artifact. Alternatively, it is possible that the -2c-containing transcript is of very low abundance, and therefore not detectable by RNAse protection. It would in future experiments be advisable to design a set of new primers to further assess the presence of -2c-containing exons both by PCR and 5'RACE. PCR is known to give rise to artifacts, but it remains puzzling as to

why it would specifically (be it at very low frequency) give rise to 5'RACE products that, when blasted against the entire mouse genome, so specifically align with genomic sequence just upstream exon -2b. That this could happen by chance is hard to imagine. One possibility is that the -2c-containing transcript has a complex secondary/tertiary structure which makes it hard to amplify by 5'RACE and/or PCR. In retrospect, further studies into this exon -2c are certainly warranted, as we did not manage to find leukocyte-specific promoter elements upstream of -2b. This might be due to other reasons, one for example that transient reporter assays are not always very physiological, but it is just possible that the region upstream of -2c is the real transcription start for leukocyte-specific p110δ mRNAs.

Determination of absolute levels of p110δ mRNA species by real time PCR

Two sets of transcript-specific mouse and human primers were designed for real time PCR experiments in an attempt to find a correlation between expression levels of p110δ transcripts 2 and 3, and the p110δ protein levels in the cells that contained these transcripts. We found that the cell lines that express highest levels of p110δ protein express high levels of transcript 2 as well as high levels of transcript 3 (Fig. 5-20). The melanoma cell line tested (B16-BL6) does not express transcript 3 but expresses high levels of transcript 2, accounting for the higher p110δ protein expression compared to other non-leukocyte cell lines. These data also suggest that the expression of transcript 3 in leukocyte cell lines is not solely responsible for the p110δ protein found in this cell type.

8.4 *In silico* promoter analysis of the mouse and human p110δ gene

There are numerous different types of software available for the detection of putative promoter sites, CpG islands, transcription factor binding sites and first exons, all of which can help towards the identification of the actual transcription start site and the promoter and other elements involved in the regulation of the gene of interest. Due to the degenerate nature of the binding of transcription factors to transcription factor binding sites, identification of functional sites is not straightforward. Genome-wide

analysis has revealed that elements in non-coding areas of DNA, which are conserved between species, tend to be involved in the regulation of the gene. Therefore homology searches across species helps to identify potential regulatory elements. We therefore carried out comparisons between human and mouse genomic DNA sequences surrounding the p110δ exons, assuming that the regulation of p110δ expression in both species was similar. The currently available evidence from tissue distribution studies of human and mouse p110δ suggest that this might indeed be the case. A complication in the studies comparing human and mouse p110δ genes are the large differences in the sizes of the introns surrounding the 5' non-coding exons (Figure 8-2).

Alignment software, transcription factor binding site identification software, two CpG island identification tools and two putative promoter identification tools were chosen to analyse the potential regulatory regions of the human and mouse p110δ gene. In contrast to the high sequence homology between the coding region of the mouse and human p110δ gene (14 kb and 18.5 kb, respectively) which allowed straightforward alignment, the alignment of the 5'UTR was not quite so straightforward, due to the large size differences in mouse and human (5'UTR is 46 kb and 87.3 kb in mouse and human, respectively). When the human and mouse sequences were submitted for alignment, large areas of highly homologous sequence were identified, but which turned out to be not conserved in position in the p110δ genome. For example a 125 bp sequence that shares 78% homology between the mouse and human gene is located 4 kb *up*stream of mouse exon -2b but it is located 10 kb *down*stream of exon -2b in the human p110δ 5'UTR. It is very well possible that this region is involved in one-way or another in the regulation of p110δ gene but most likely not in the initiation of transcription at the promoter, due to its position in the gene.

The other six highly homologous regions identified in the 5'UTR of the mouse and human p110δ gene lie between exon -2a and exon -1 (Figure 6-3). In fact, this area is the least interesting, as it are most likely sequences 5' of exon -2a, and especially sequences 5' of exon -2b (leukocyte-specific) that are most likely critical for transcription of transcripts 2 and 3. Five out of the six homology areas are significant distances 5' from exon -1 (24 kb, 22 kb, 21 kb, 20.5 kb and 13 kb) therefore may be

involved in regulation of p110δ expression but again are unlikely to be directly involved in p110δ gene promoter activity. The closest large region of homology in the 5'UTR of the human and mouse p110δ gene is 2 kb upstream of exon -1. When tested in a reporter assay, however, this region did not yield significant promoter activity (discussed in more detail in section 8.5).

In an attempt to resolve the issue of the size difference between exons and to identify smaller areas of homology, we focused our attention to 250 bp upstream of each exon and submitted this sequence to rVISTA for detailed transcription factor binding site analysis (note that such analyses cannot be carried out on large DNA sequences, as they yield vast amounts of data, given that transcription factor binding sites are only 6 to 8 bp on average). In summary, our findings were as follows:

- In the ~350 bp surrounding and including **exon -1**, a 98 bp region with 73.5% homology between human and mouse was identified. rVISTA identified a homologous TATA signal 29 bp and 40 bp upstream of mouse and human exon -1, respectively. The conserved position of the TATA box is -25 bp from the transcriptional start site making these possible but not definite promoter elements (Figure 6-5).

- A 138 bp region with 75.4% homology was identified within the ~300-400 bp aligned surrounding **exon -2a.** The homologous TATA signal identified upstream of exon -2a is 105 bp and 217 bp from the transcription start site of the mouse and human exon, respectively, therefore unlikely to be functional (Figure 6-5B).

- When similarly stringent parameters of analysis were applied to the mouse and human exon -2b no homology was found, even when 250 bp of intron sequence upstream of the exon was added to the alignment (Figure 6-4).

We next analysed smaller areas upstream of the mouse and human p110δ 5' untranslated exons using less stringent parameters. This revealed numerous potential leukocyte-specific transcription factor-binding sites upstream of both the mouse and

human 5'UTR exons (for a summary see Table 6-1). However, some of these sites were not conserved between human and mouse.

There are several CpG clusters called "CpG islands" in vertebrate genomes, and they thought to be around promoter region. The FIRST EF software revealed CpG-related promoters upstream of all three human and mouse 5' untranslated exons, but unfortunately not upstream of mouse exon -2b. GRAIL and Softberry software confirmed the presence of CpG islands within all of these FIRST EF identified putative promoter regions (for a summary see Figure 6-6).

8.5 Cell-based p110δ promoter analysis using reporter assays

Bioinformatic analysis of potential gene regulatory sequences must be confirmed by functional analysis. The information collated from bioinformatics analysis was next used as a guide in designing the reporter constructs for a more detailed *in vitro* analysis of p110δ gene expression using reporter assays.

A 2 kb sequence upstream of each of the 5' untranslated exons in the mouse p110δ gene was amplified from mouse genomic DNA and cloned into a luciferase reporter vector. Deletion constructs were created from these larger PCR fragments in an attempt to identify the presence of regulatory elements in the mouse p110δ gene.

As discussed above, transcript 3 (incorporating exon -2b) is only expressed in leukocyte cell lines. Transcript 2 (incorporating exon -2a) is present at similar levels in leukocyte and non-leukocyte cell lines that express intermediate levels of p110δ (such as melanoma cells) but is higher than in non-leukocytes such as NIH3T3 which express barely detectable levels of p110δ mRNA and protein (Fig. 5-20). We therefore hoped, when carrying out reporter studies in leukocytes and NIH 3T3 cells, to identify regulatory elements which only worked in the leukocytes, and not in the NIH 3T3 cells.

Functional analysis of DNA sequences upstream of exon –2b (analysis of putative promoter for transcript 3)

No putative promoter sites or CpG islands were identified upstream of exon -2b in the mouse p110δ gene. We nevertheless created 6 different sized reporter constructs containing DNA fragments, ranging from 2 kb to 58 bp, from the genomic DNA upstream of exon -2b. These constructs were then transiently transfected in NIH3T3, a non-leukocyte cell line, and A20, a leukocyte cell line. The 2 kb DNA fragment incorporated homologous region A (Figure 6-3) which aligned with a region downstream of human exon -2b. Reporter activity recorded from the 2 kb DNA fragment was very low (Figure 7-4) in both A20 and NIH3T3 cells. Surprisingly, the activity increased when 800 bp of DNA was removed from the 5' end. However, the activity recorded for this 1195 bp fragment was still only just higher than the activity detected using the promoterless control. In fact, the reporter activity recorded from all six experimental samples was never 1.5 times above the activity recorded from the promoterless control in either cell line. The transfections had been effective, based on the activity of the co-transfected *renilla* luciferase reporter vector. In addition, tissue-specific (i.e. leukocyte-specific) reporter activity was found in p110γ gene promoter constructs that were used as positive controls (174). From these results it is difficult to draw firm conclusions about the involvement of any of the predicted transcription factor binding sites in the regulation of p110δ gene expression.

We next carried out stable transfections of three of the mouse exon -2b constructs (Figure 7-5), and invariably found higher promoter activity in NIH3T3 cells than in A20 cells. This was surprising, as we had expected the opposite, as results obtained thus far suggest that transcript 3 (incorporating exon -2b) is not expressed in non-leukocyte cell lines, on the basis of which we did not expect to find promoter activity in fibroblasts. A leukocyte-specific promoter from Vav, used as a positive control, showed a 50-fold increase in promoter activity in the A20 cells compared to the fibroblasts, as expected.

Functional analysis of DNA sequences upstream of exon –2a (analysis of putative promoter for transcript 2)

Seven constructs incorporating region upstream of exon -2a of the mouse p110δ gene gave rise to a 2- to 3-fold increase in reporter activity upon transient transfection, in both A20 and NIH 3T3 cells. Remarkably, even when 86% of the putative (predicted) promoter was removed, leaving only a 36 bp DNA fragment, a 2- and 2.8 -fold increase in reporter activity was still observed in NIH 3T3 and A20, respectively. There was no *in silico* evidence to suggest that the 36 bp region immediately upstream of exon -2a in the mouse p110δ gene contains a regulatory element, therefore without doing further deletion studies it is not possible to determine the functionality of this region.

We next stably transfected three exon -2a constructs in NIH 3T3 and A20 cells, and found similar levels of promoter activity in both cell lines. Real time PCR analysis demonstrated higher expression levels of transcript 2 (incorporating exon -2a) in A20 cells than in NIH 3T3 cells (Figure 5-20). It is not clear why this was not reflected in the results of our transient or stable reporter assays. One explanation could be the presence of tissue-specific regulatory elements more than 2 kb upstream of exon -2a. Sequences further than 2 kb 5' of exon -2a were not included in the reporter assay. There was a significant drop (6-fold) in reporter activity between the 2 kb construct and the 1.5 kb construct in both the stably transfected cell lines. *In silico* analysis of this 500 bp region did not identify any putative regulatory elements, making it hard to account for this drop in promoter activity.

Functional analysis of DNA sequences upstream of exon –1 (analysis of putative promoter for transcript 1)

In transient transfections, reporter constructs containing DNA from the genomic region upstream of exon -1 had higher promoter activity than the genomic region upstream of either of the other two exons, -2a and -2b. Indeed, there was over a 3-fold increase in promoter activity from nearly all of the constructs. Two significant results were obtained from this set of experiments:

1) The largest putative promoter constructs were significantly more active in NIH 3T3 than in A20 cells. The reporter activity in the B cells increased to

similar levels as those seen in fibroblasts upon the deletion of 315 bp from the 5' end of the p110δ fragment (Figure 7-12). One possible explanation for this is that a tissue-specific repressor element was present in the 315 bp of DNA that was removed.

2) Compared to the promoterless control, the putative promoter construct containing the 185 bp of genomic DNA immediately upstream of exon -1, gave rise to a 12-fold increase in reporter activity in the A20 cell line in contrast to a 3-fold increase in the NIH 3T3 cells. The increase in reporter activity from the 185 bp putative promoter construct compared to the 539 bp putative promoter construct (Figure 7-12) suggests that a region involved in silencing was removed during the deletion process. Therefore there may be a tissue-specific regulatory region within this 354 bp region. Further analysis would need to be carried out to confirm this hypothesis.

A number of breast cell lines were identified as expressing transcript 1 (incorporating exon -1 alone) (Table 5-3). No breast cell lines were used in the reporter assays therefore the tissue-specific reporter activity demonstrated above from certain p110δ constructs was unexpected and we were unable to come up with a logical explanation.

Functional analysis of DNA sequences upstream of exon 1

The 3 transcripts identified by 5'RACE all incorporated exon -1 and exon 1 (Fig. 8-1). None of the transcripts appear to start at exon 1, suggesting that there is no promoter immediately upstream of exon 1. However, it does not rule out the possibility of the presence of regulatory elements within this region.

Very remarkably, the reporter activity recorded from transient transfections of constructs containing putative regulatory elements upstream of exon 1, in both the leukocyte and fibroblast cells, was significantly higher than activity recorded in the assays using sequences flanking exons -2b, -2a and -1, described above. A region of 192 bp (-64 to -254) gave rise to an over 20-fold increase in reporter activity compared to the promoterless control, in both the NIH 3T3 and A20 cells (Figure 7-15). The activity decreased to just over 2-fold when a further 6 bp were deleted from the fragment. The lack of *in silico* evidence of a putative promoter in this region

suggests that it may contain an enhancer. Further analysis of this 6 bp sequence failed to identify a binding site for a transcription factor. One possible hypothesis is that this small region, in partnership with a region further downstream, may be required for the 20-fold increase in reporter activity.

A construct with a further 13 bp deleted yielded no reporter activity suggesting that the 20 bp, -44 to -54 from the ATG translation start site, are in some way involved in the regulation of p110δ gene expression. A number of transcription factor binding sites were identified in the 62 bp region upstream of the translation start site of exon 1. Constructs created with mutations in some of these sites failed to identify any particular elements that may be involved in the regulation of the p110δ protein expression, in that none of the constructs made lost the capacity to drive reporter gene expression (data not shown).

Functional analysis of DNA sequences upstream of exons -2b, -2a, -1 and 1 in cells other than NIH 3T3 and A20

The largest and smallest DNA construct, created from genomic DNA upstream of each of the four p110δ exons studied (exon -2b, exon -2a, exon -1 and exon 1), were transiently transfected into another mouse leukocyte cell line, WEHI 231, and two different mouse epithelial cell lines, CT26 and LLC. The results obtained from the WEHI 231 cell line reproduced the results obtained in the A20 cell line. Similarly the CT2 and LLC results followed suit, reproducing the data obtained from the NIH3T3 cells (data not shown). This confirms that the observations are reproducible between cell types.

Conclusion on reporter analysis

On the basis of our results, it is not possible to delineate specific sites in the genomic DNA which can account for the tissue-specific regulation of the p110δ gene. It is not yet clear whether the expression of transcript 3 (incorporating exon -2b) is responsible for the increase in p110δ expression in leukocytes, or whether it is as a result of leukocyte-specific transcription factors acting further downstream, or whether it is a combination of the two factors. A third possibility is silencing factors acting in non-leukocyte cell lines. At this stage none of these hypotheses can be ruled out.

One conclusion that can be drawn from our work is that reporter assays can be potentially misleading. Unless there is evidence of a significant increase or decrease in promoter activity from particular constructs, for example the Vav reporter constructs, no conclusions can be drawn. Transient transfections do not represent the real *in vivo* chromatin state in which the gene normally resides. Likewise, stable transfections may lead to integration into an area which does not reflect the normal regulation of the gene, as the reporter construct may 'land' in an over-expressed or tightly repressed area of the genome.

The next step forward would be to try and determine the chromatin state of the area surrounding the p110δ gene using DNase hypersensitivity studies (described in section 8.7). Small areas surrounding the p110δ gene in leukocytes may be less tightly bound with histones therefore allowing transcription factors to bind freely. In fibroblasts this may not be the case, the chromatin may be more tightly packed therefore preventing such a high rate of transcription.

8.6 Conclusions and future work

We have shown that the p110δ gene is regulated at or before the level of transcription and that a number of untranslated exons 5' of the translation start site are expressed in a tissue-specific manner. Reporter assays failed to identify one particular area of tissue specific promoter activity suggesting that other regulatory factors may be involved *in vivo*. Experiments to try to identify whether the chromatin structure surrounding the p110δ gene play a role in its regulation used DNA demethylating agents and reagents that inhibit histone deacetylation. DNA methylation and histone acetylation are mechanisms that control gene expression by dictating transitions between transcriptionally active or transcriptionally silent chromatin states. A 2-fold increase in p110δ mRNA expression was observed upon treatment with TSA but this was not reflected in an equal increase of p110δ expression at the protein level. This was not a fully conclusive result obtained from only one set of experiments therefore I still believe that it is possible that *in vivo* CpG hypermethylation and/or a strongly regulated tight chromatin structure may play a role in the regulation of the p110δ

gene. Without further experiments, such as DNase hypersensitivity, we cannot rule this out as a reason for the inconclusive results obtained from the reporter assays.

At the onset of my PhD project there was no database evidence for the existence of the 5'-untranslated exons discussed in this thesis. Mass sequencing projects have slowly uncovered evidence for these upstream exons. To date, p110δ is the only class I PI3K with 3 untranslated exons 5' of the translation start site (Figure 5-21). Information from databases suggests that there is only one upstream-untranslated exon belonging to p110α and p110γ and two belonging to p110β.

If I had more time in the laboratory the next experiment I would undertake to unravel the regulation of p110δ gene expression would be to

(1) put further effort in finding out whether exon -2c exists (discussed in paragraph 8.3 above)
(2) focus the attention of promoter elements upstream of exon -2b, as this is leukocyte-specific, and carry out more detailed functional analysis
(3) carry out DNAse hypersensitivity studies. This approach is further discussed below.

Mapping DNase-hypersensitive sites (HSS) has been used to identify the precise location of many different regulatory elements in genes. Promoters, enhancers, suppressors, insulators, and locus control regions all have been shown to be associated with DNase HSS (24). This approach has the advantage of taking chromatin context into account by digesting nucleosome-free regions of the genome, allowing for identification of both ubiquitous and tissue-specific regulatory elements.

A locus control region (LCR) is a DNA regulatory element. It is widely believed that LCRs function by providing an open chromatin environment. This chromatin-opening ability is thought to be inherently tissue-specific in nature. It has been observed that the activity of an LCR is restricted to the tissues in which its locus of origin is normally expressed. LCRs are thought to be important in the determination of tissue-specific gene expression in the context of chromatin. Classical transcriptional regulatory elements (enhancers/silencers), with their associated nuclear factors, have also been implicated in tissue-specific gene regulation, as some of them are cell type-

specific in their activity. The role of these elements is thought to be the modulation of RNA polymerase activity. Their action is thought to be subsequent to LCR-mediated chromatin opening, which allows nuclear factor access to their DNA-binding sites (24,215).

In practical terms, these experiments briefly go as follows (Fig. 8-1). Nuclei are isolated from the cell lines of interest (for example a leukocyte and a non-leukocyte in our case) and digested with a low concentration of DNase I. DNA which is inaccessible, as a consequence of being wrapped up in a tight chromatin structure, is not accessible for the DNase and remains uncut. Chromatin structure surrounding genes may differ between cell lines. So, for example, if leukocytes have an area of genomic DNA that is accessible to DNase, and the non-leukocytes do not, then different DNA fragments will be generated in these cell lines. The more accessible areas of DNA may be involved in the regulation of the gene. For example, they may be transcription factor binding sites or gene promoters that are only accessible in certain cell types.

The genomic DNA is cut with a restriction enzyme of choice to create a fragment of known size containing the area of interest. This could for example be a 2 kb fragment surrounding exon -2b in the case of the p110δ gene. A radioactively labelled probe is designed to be complementary to the ~200 bp sequence at the 5' end of this fragment. The digested genomic DNA is separated on an agarose gel, followed by Southern blot to hybridise the probe to the digested, separated, genomic DNA (Fig 8-4). The number and size of the radioactively labeled genomic DNA fragments that will be detected depends on the number and situation/location of the DNAse hypersensitive sites.

Due to the large size of the intron sequences in between the 5' upstream exons in the human and mouse p110δ gene, it would be advisable to analyze each exon and its surrounding intron sequence separately. Comparison of autoradiograms produced from leukocyte and non-leukocyte cell lines may reveal a tissue-specific difference in the chromatin structure of the two cell types.

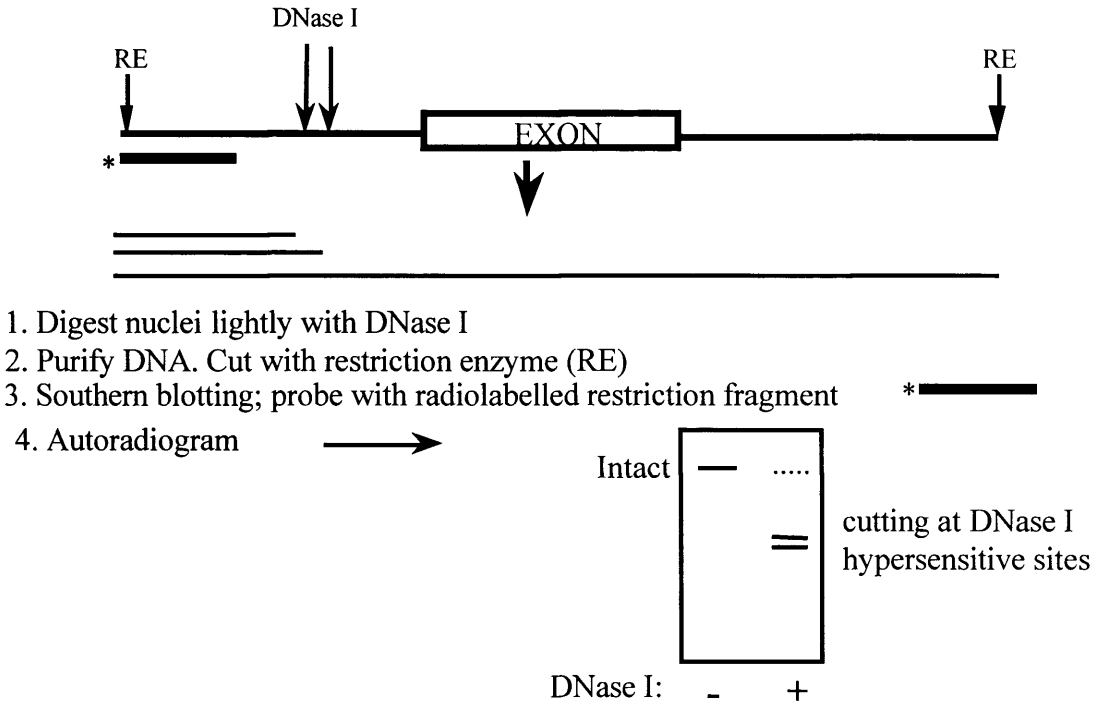

Figure 8-4 Diagrammatic representation of a DNase hypersensitivity assay

If these results yielded useful information I would attempt to determine any proteins that bound to the DNA in the DNase hypersensitive regions, initially by using the gel mobility-shift assay. Radioactively labeled DNA probes, containing the DNase HSS, are incubated with total cell lysate from the cell line of interest. The reaction is analysed on a polyacrylamide gel run under conditions in which bound protein remain with the DNA fragment. Free fragment runs to a position characteristic of its size, usually at the bottom of the gel. Fragments with proteins bound are much larger therefore migrate more slowly through the gel, and appears as a "shifted" fragment of DNA higher up in the gel.

At this stage if the experiments were still going to plan, we could consider to use the DNA as an affinity probe to isolate the proteins that bind it. This could for example be done by covalently coupling this DNA to a matrix such as sepharose and incubating it with a cellular/nuclear extract (216). Then it may be possible to identify protein-binding partners by mass spectrometry.

If different HSS were discovered in leukocyte and non-leukocyte cell lines, we may be able to conclude that the chromatin structure upstream of exon -2b in non-leukocytes is more tightly bound in histones therefore not accessible to the transcription factor that transcribe the same transcript in leukocyte cell lines.

Alternatively if the DNase hypersensitivity patterns were similar in leukocytes and non-leukocytes, the gel mobility-shift assay and mass spectrometry may shed light onto possible differences in transcription factor binding.

These are all 'what ifs' and from the work done for this thesis I now realize that things are never quite as straight forward as they first seem. To me, it appears as if it may be many years until the regulation of the p110δ gene will be fully delineated.

9 References

1. Vanhaesebroeck, B., Leevers, S. J., Ahmadi, K., Timms, J., Katso, R., Driscoll, P. C., Woscholski, R., Parker, P. J., and Waterfield, M. D. (2001) *Annu Rev Biochem* **70**, 535-602
2. Fruman, D. A., and Cantley, L. C. (2002) *Semin Immunol* **14**, 7-18
3. Koyasu, S. (2003) *Nat Immunol* **4**, 313-319
4. Okkenhaug, K., and Vanhaesebroeck, B. (2003) *Nat Rev Immunol* **3**, 317-330
5. Fruman, D. A., Meyers, R. E., and Cantley, L. C. (1998) *Annu Rev Biochem* **67**, 481-507
6. Rameh, L. E., and Cantley, L. C. (1999) *J Biol Chem* **274**, 8347-8350
7. Sawyer, C., Sturge, J., Bennett, D. C., O'Hare, M. J., Allen, W. E., Bain, J., Jones, G. E., and Vanhaesebroeck, B. (2003) *Cancer Res* **63**, 1667-1675
8. Vanhaesebroeck, B., Welham, M. J., Kotani, K., Stein, R., Warne, P. H., Zvelebil, M. J., Higashi, K., Volinia, S., Downward, J., and Waterfield, M. D. (1997) *Proc Natl Acad Sci U S A* **94**, 4330-4335
9. Vanhaesebroeck, B., Jones, G. E., Allen, W. E., Zicha, D., Hooshmand-Rad, R., Sawyer, C., Wells, C., Waterfield, M. D., and Ridley, A. J. (1999) *Nat Cell Biol* **1**, 69-71
10. Lodish H, B. A., Zipursky S, Matsudaria P, Baltimore D and Darnell J. (2000) *Molecular Cell Biology Fourth Edition*, Fourth Edition Ed., W.H Freeman and Company
11. Alberts B, B. D., Lewis J, Raff M, Roberts K, Watson J D. (1994) *Molecular Biology of the Cell*, Third Edition Ed., Garland Publishing Inc.
12. Devlin, T. M. (1997) *Textbook of Biochemistry with Clinical Correlations Fourth Edition*, Fourth Edition Ed., Whiley Liss
13. Kornberg, R. D. (1974) *Science* **184**, 868-871
14. Thomas, J. O., and Kornberg, R. D. (1975) *Proc Natl Acad Sci U S A* **72**, 2626-2630
15. Kornberg, R. D., and Lorch, Y. (1999) *Cell* **98**, 285-294
16. Finch, J. T., Lutter, L. C., Rhodes, D., Brown, R. S., Rushton, B., Levitt, M., and Klug, A. (1977) *Nature* **269**, 29-36
17. Grunstein, M. (1990) *Annu Rev Cell Biol* **6**, 643-678
18. Kornberg, R. D., and Lorch, Y. (1992) *Annu Rev Cell Biol* **8**, 563-587
19. Grunstein, M. (1990) *Trends Genet* **6**, 395-400
20. Wu, J., and Grunstein, M. (2000) *Trends Biochem Sci* **25**, 619-623
21. Almer, A., Rudolph, H., Hinnen, A., and Horz, W. (1986) *Embo J* **5**, 2689-2696
22. Fascher, K. D., Schmitz, J., and Horz, W. (1990) *Embo J* **9**, 2523-2528
23. Hirschhorn, J. N., Brown, S. A., Clark, C. D., and Winston, F. (1992) *Genes Dev* **6**, 2288-2298
24. Gross, D. S., and Garrard, W. T. (1988) *Annu Rev Biochem* **57**, 159-197
25. Vignali, M., Hassan, A. H., Neely, K. E., and Workman, J. L. (2000) *Mol Cell Biol* **20**, 1899-1910
26. Kornberg, R. D., and Lorch, Y. (1999) *Curr Opin Genet Dev* **9**, 148-151
27. Sudarsanam, P., Cao, Y., Wu, L., Laurent, B. C., and Winston, F. (1999) *Embo J* **18**, 3101-3106

28. Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., and Young, R. A. (1998) *Cell* **95**, 717-728
29. Kornberg, R. D., and Lorch, Y. (1991) *Cell* **67**, 833-836
30. Utley, R. T., Cote, J., Owen-Hughes, T., and Workman, J. L. (1997) *J Biol Chem* **272**, 12642-12649
31. Cardinaux, J. R., Chapel, S., and Wahli, W. (1994) *J Biol Chem* **269**, 32947-32956
32. Watson J, W. J., Gilman M and Zoller M. (1996) *Recombinant DNA Second Edition*, W.H Freeman and Company
33. Keshet, I., Lieman-Hurwitz, J., and Cedar, H. (1986) *Cell* **44**, 535-543
34. Tazi, J., and Bird, A. (1990) *Cell* **60**, 909-920
35. Lewis, J. D., Meehan, R. R., Henzel, W. J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. (1992) *Cell* **69**, 905-914
36. Meehan, R. R., Lewis, J. D., and Bird, A. P. (1992) *Nucleic Acids Res* **20**, 5085-5092
37. Nan, X., Campoy, F. J., and Bird, A. (1997) *Cell* **88**, 471-481
38. Goto, T., and Monk, M. (1998) *Microbiol Mol Biol Rev* **62**, 362-378
39. Machwe, A., Orren, D. K., and Bohr, V. A. (2000) *Faseb J* **14**, 1715-1724
40. Jones, P. A., and Baylin, S. B. (2002) *Nat Rev Genet* **3**, 415-428
41. Bestor, T. H. (1998) *Nature* **393**, 311-312
42. Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., and Bird, A. (1998) *Nature* **393**, 386-389
43. Hsieh, C. L. (1994) *Mol Cell Biol* **14**, 5487-5494
44. Chen, Z. J., and Pikaard, C. S. (1997) *Genes Dev* **11**, 2124-2136
45. Wade, P. A., Pruss, D., and Wolffe, A. P. (1997) *Trends Biochem Sci* **22**, 128-132
46. Lewin, B. (1997) *Genes VI*, Oxford University Press
47. Strahle, U., Schmidt, A., Kelsey, G., Stewart, A. F., Cole, T. J., Schmid, W., and Schutz, G. (1992) *Proc Natl Acad Sci U S A* **89**, 6731-6735
48. Kos, M., O'Brien, S., Flouriot, G., and Gannon, F. (2000) *FEBS Lett* **477**, 15-20
49. Leid, M., Kastner, P., and Chambon, P. (1992) *Trends Biochem Sci* **17**, 427-433
50. Lazar, M. A. (1993) *Endocr Rev* **14**, 184-193
51. Liu, M., Whetstine, J. R., Payton, S. G., Ge, Y., Flatley, R. M., and Matherly, L. H. (2004) *Biochem J* **383**, 249-257
52. Streb, J. W., Kitchen, C. M., Gelman, I. H., and Miano, J. M. (2004) *J Biol Chem* **279**, 56014-56023
53. Falkenberg, V. R., Alvarez, K., Roman, C., and Fregien, N. (2003) *Glycobiology* **13**, 411-418
54. Sentenac, A. (1985) *CRC Crit Rev Biochem* **18**, 31-90
55. Hampsey, M. (1998) *Microbiol Mol Biol Rev* **62**, 465-503
56. Breathnach, R., and Chambon, P. (1981) *Annu Rev Biochem* **50**, 349-383
57. Azizkhan, J. C., Jensen, D. E., Pierce, A. J., and Wade, M. (1993) *Crit Rev Eukaryot Gene Expr* **3**, 229-254
58. Widen, S. G., Kedar, P., and Wilson, S. H. (1988) *J Biol Chem* **263**, 16992-16998
59. Weis, L., and Reinberg, D. (1992) *Faseb J* **6**, 3300-3309

60. Martini, G., Toniolo, D., Vulliamy, T., Luzzatto, L., Dono, R., Viglietto, G., Paonessa, G., D'Urso, M., and Persico, M. G. (1986) *Embo J* **5**, 1849-1855
61. Lin, Y., Ince, T. A., and Scotto, K. W. (2001) *Biochemistry* **40**, 12959-12966
62. Ince, T. A., and Scotto, K. W. (1995) *Gene* **156**, 287-290
63. Ince, T. A., and Scotto, K. W. (1995) *J Biol Chem* **270**, 30249-30252
64. Conaway, R. C., and Conaway, J. W. (1993) *Annu Rev Biochem* **62**, 161-190
65. Sawadogo, M., and Roeder, R. G. (1985) *Proc Natl Acad Sci U S A* **82**, 4394-4398
66. Zawel, L., and Reinberg, D. (1993) *Prog Nucleic Acid Res Mol Biol* **44**, 67-108
67. Buratowski, S., Hahn, S., Guarente, L., and Sharp, P. A. (1989) *Cell* **56**, 549-561
68. Robert, F., Forget, D., Li, J., Greenblatt, J., and Coulombe, B. (1996) *J Biol Chem* **271**, 8517-8520
69. Parvin, J. D., and Sharp, P. A. (1993) *Cell* **73**, 533-540
70. Svejstrup, J. Q., Wang, Z., Feaver, W. J., Wu, X., Bushnell, D. A., Donahue, T. F., Friedberg, E. C., and Kornberg, R. D. (1995) *Cell* **80**, 21-28
71. Koleske, A. J., and Young, R. A. (1995) *Trends Biochem Sci* **20**, 113-116
72. Sopta, M., Burton, Z. F., and Greenblatt, J. (1989) *Nature* **341**, 410-414
73. Zawel, L., Kumar, K. P., and Reinberg, D. (1995) *Genes Dev* **9**, 1479-1490
74. Workman, J. L., and Roeder, R. G. (1987) *Cell* **51**, 613-622
75. Dikstein, R. (2001) in *Cell communication and Signal transduction*
76. Nikolov, D. B., and Burley, S. K. (1994) *Nat Struct Biol* **1**, 621-637
77. Li, R., Knight, J. D., Jackson, S. P., Tjian, R., and Botchan, M. R. (1991) *Cell* **65**, 493-505
78. Dillon, N., and Grosveld, F. (1993) *Trends Genet* **9**, 134-137
79. Walters, M. C., Fiering, S., Eidemiller, J., Magis, W., Groudine, M., and Martin, D. I. (1995) *Proc Natl Acad Sci U S A* **92**, 7125-7129
80. Walters, M. C., Magis, W., Fiering, S., Eidemiller, J., Scalzo, D., Groudine, M., and Martin, D. I. (1996) *Genes Dev* **10**, 185-195
81. McKnight, S. L., Gavis, E. R., Kingsbury, R., and Axel, R. (1981) *Cell* **25**, 385-398
82. Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. (2000) *Science* **288**, 136-140
83. Shatkin, A. J., and Manley, J. L. (2000) *Nat Struct Biol* **7**, 838-842
84. Parent, A. B., I. Bougie,I and Bisaillon,M. (2004) *Journal of Biological Sciences* **4**, 624-627
85. Varani, G. (1997) *Structure* **5**, 855-858
86. McCracken, S., Fong, N., Rosonina, E., Yankulov, K., Brothers, G., Siderovski, D., Hessel, A., Foster, S., Shuman, S., and Bentley, D. L. (1997) *Genes Dev* **11**, 3306-3318
87. Sachs, A. B., Sarnow, P., and Hentze, M. W. (1997) *Cell* **89**, 831-838
88. Zhao, J., Hyman, L., and Moore, C. (1999) *Microbiol Mol Biol Rev* **63**, 405-445
89. Kramer, A. (1996) *Annu Rev Biochem* **65**, 367-409
90. Moore MJ, Q. C., Sharp PA. (1993) in *The RNA World* (Gesteland R, A. J., ed), pp. 303-357, Cold Spring Harbour Laboratory Press
91. Brow, D. A. (2002) *Annu Rev Genet* **36**, 333-360
92. Rappsilber, J., Ryder, U., Lamond, A. I., and Mann, M. (2002) *Genome Res* **12**, 1231-1245

93. Hartmuth, K., Urlaub, H., Vornlocher, H. P., Will, C. L., Gentzel, M., Wilm, M., and Luhrmann, R. (2002) *Proc Natl Acad Sci U S A* **99**, 16719-16724
94. Jurica, M. S., Licklider, L. J., Gygi, S. R., Grigorieff, N., and Moore, M. J. (2002) *Rna* **8**, 426-439
95. Ladd, A. N., and Cooper, T. A. (2002) *Genome Biol* **3**, reviews0008
96. Zahler, A. M., Neugebauer, K. M., Lane, W. S., and Roth, M. B. (1993) *Science* **260**, 219-222
97. Cartegni, L., Chew, S. L., and Krainer, A. R. (2002) *Nat Rev Genet* **3**, 285-298
98. Lopez, A. J. (1998) *Annu Rev Genet* **32**, 279-305
99. Moore, M. J. (2002) *Science* **298**, 370-371
100. Wagner, E., and Lykke-Andersen, J. (2002) *J Cell Sci* **115**, 3033-3038
101. Kastner, P., Krust, A., Turcotte, B., Stropp, U., Tora, L., Gronemeyer, H., and Chambon, P. (1990) *Embo J* **9**, 1603-1614
102. Keaveney, M., Klug, J., Dawson, M. T., Nestor, P. V., Neilan, J. G., Forde, R. C., and Gannon, F. (1991) *J Mol Endocrinol* **6**, 111-115
103. Zennaro, M. C., Keightley, M. C., Kotelevtsev, Y., Conway, G. S., Soubrier, F., and Fuller, P. J. (1995) *J Biol Chem* **270**, 21016-21020
104. Castren, M., and Damm, K. (1993) *J Neuroendocrinol* **5**, 461-466
105. Christakos, S., Raval-Pandya, M., Wernyj, R. P., and Yang, W. (1996) *Biochem J* **316 (Pt 2)**, 361-371
106. Crofts, L. A., Hancock, M. S., Morrison, N. A., and Eisman, J. A. (1998) *Proc Natl Acad Sci U S A* **95**, 10529-10534
107. Voronova, A. F., Adler, H. T., and Sefton, B. M. (1987) *Mol Cell Biol* **7**, 4407-4413
108. Wildin, R. S., Wang, H. U., Forbush, K. A., and Perlmutter, R. M. (1995) *J Immunol* **155**, 1286-1295
109. Adler, H. T., Reynolds, P. J., Kelley, C. M., and Sefton, B. M. (1988) *J Virol* **62**, 4113-4122
110. Garvin, A. M., Pawar, S., Marth, J. D., and Perlmutter, R. M. (1988) *Mol Cell Biol* **8**, 3058-3064
111. Goodnight, J. A., Mischak, H., Kolch, W., and Mushinski, J. F. (1995) *J Biol Chem* **270**, 9991-10001
112. Koizumi, H., Kohno, Y., Osada, S., Ohno, S., Ohkawara, A., and Kuroki, T. (1993) *J Invest Dermatol* **101**, 858-863
113. Quan, T., and Fisher, G. J. (1999) *J Biol Chem* **274**, 28566-28574
114. Nudel, U. a. Y. D. (2004) in *Faculty of Biology*, pp. 116-118, Department of Molecular Cell Biology
115. Cox, G. F., and Kunkel, L. M. (1997) *Curr Opin Cardiol* **12**, 329-343
116. Carr, C., Fischbach, G. D., and Cohen, J. B. (1989) *J Cell Biol* **109**, 1753-1764
117. Kramarcy, N. R., Vidal, A., Froehner, S. C., and Sealock, R. (1994) *J Biol Chem* **269**, 2870-2876
118. Holzfeind, P. J., Ambrose, H. J., Newey, S. E., Nawrotzki, R. A., Blake, D. J., and Davies, K. E. (1999) *J Biol Chem* **274**, 6250-6258
119. Nudel, U., Zuk, D., Einat, P., Zeelon, E., Levy, Z., Neuman, S., and Yaffe, D. (1989) *Nature* **337**, 76-78
120. Gorecki, D. C., Monaco, A. P., Derry, J. M., Walker, A. P., Barnard, E. A., and Barnard, P. J. (1992) *Hum Mol Genet* **1**, 505-510
121. Thomas, H., Jaschkowitz, K., Bulman, M., Frayling, T. M., Mitchell, S. M., Roosen, S., Lingott-Frieg, A., Tack, C. J., Ellard, S., Ryffel, G. U., and Hattersley, A. T. (2001) *Hum Mol Genet* **10**, 2089-2097

122. Yamagata, K., Furuta, H., Oda, N., Kaisaki, P. J., Menzel, S., Cox, N. J., Fajans, S. S., Signorini, S., Stoffel, M., and Bell, G. I. (1996) *Nature* **384**, 458-460
123. Adams, J. M., Houston, H., Allen, J., Lints, T., and Harvey, R. (1992) *Oncogene* **7**, 611-618
124. Ogilvy, S., Elefanty, A. G., Visvader, J., Bath, M. L., Harris, A. W., and Adams, J. M. (1998) *Blood* **91**, 419-430
125. Mercatante, D., and Kole, R. (2000) *Pharmacol Ther* **85**, 237-243
126. Suwanmanee, T., Sierakowska, H., Lacerra, G., Svasti, S., Kirby, S., Walsh, C. E., Fucharoen, S., and Kole, R. (2002) *Mol Pharmacol* **62**, 545-553
127. De Angelis, F. G., Sthandier, O., Berarducci, B., Toso, S., Galluzzi, G., Ricci, E., Cossu, G., and Bozzoni, I. (2002) *Proc Natl Acad Sci U S A* **99**, 9456-9461
128. Rossoll, W., Kroning, A. K., Ohndorf, U. M., Steegborn, C., Jablonka, S., and Sendtner, M. (2002) *Hum Mol Genet* **11**, 93-105
129. Naimi, B., Latil, A., Fournier, G., Mangin, P., Cussenot, O., and Berthon, P. (2002) *Prostate* **52**, 245-252
130. De Moerlooze, L., Spencer-Dene, B., Revest, J., Hajihosseini, M., Rosewell, I., and Dickson, C. (2000) *Development* **127**, 483-492
131. Slusher, L. B., Gillman, E. C., Martin, N. C., and Hopper, A. K. (1991) *Proc Natl Acad Sci U S A* **88**, 9789-9793
132. Fraser, A. G., Kamath, R. S., Zipperlen, P., Martinez-Campos, M., Sohrmann, M., and Ahringer, J. (2000) *Nature* **408**, 325-330
133. Zamore, P. D. (2001) *Nat Struct Biol* **8**, 746-750
134. Sharp, P. A. (2001) *Genes Dev* **15**, 485-490
135. Tang, G. (2005) *Trends Biochem Sci* **30**, 106-114
136. Kioussis, D. (2005) *Nature* **435**, 579-580
137. Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R., and Flavell, R. A. (2005) *Nature* **435**, 637-645
138. Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S., and Gingeras, T. R. (2005) *Science* **308**, 1149-1154
139. Goueli, S. (2000) in *High Throughput Systems for drug discovery* (corp., P., ed), pp. 24-27
140. Ullrich, A., and Schlessinger, J. (1990) *Cell* **61**, 203-212
141. Superti-Furga, G., and Courtneidge, S. A. (1995) *Bioessays* **17**, 321-330
142. Garbers, D. L. (1989) *J Biol Chem* **264**, 9103-9106
143. Pawson, T. (1995) *Nature* **373**, 573-580
144. Wymann, M. P., and Pirola, L. (1998) *Biochim Biophys Acta* **1436**, 127-150
145. Wymann, M. P. M., R. (2005) *Curr Opin Cell Biol* **17**, 141-149
146. Drees B, G. M., Christian Rommel and Glenn Prestwich. (2004) in *Expert Opin. Ther. Patents* Vol. 5, pp. 703-732
147. Ali, K., Bilancio, A., Thomas, M., Pearce, W., Gilfillan, A. M., Tkaczyk, C., Kuehn, N., Gray, A., Giddings, J., Peskett, E., Fox, R., Bruce, I., Walker, C., Sawyer, C., Okkenhaug, K., Finan, P., and Vanhaesebroeck, B. (2004) *Nature* **431**, 1007-1011
148. Vanhaesebroeck, B., and Waterfield, M. D. (1999) *Exp Cell Res* **253**, 239-254
149. Vanhaesebroeck, B., Leevers, S. J., Panayotou, G., and Waterfield, M. D. (1997) *Trends Biochem Sci* **22**, 267-272

150. Roche, S., Koegl, M., and Courtneidge, S. A. (1994) *Proc Natl Acad Sci U S A* **91**, 9185-9189
151. Hill, K. M., Huang,Y, Yip,S, Yu,J, Segall,J.E, and Backer,J.M. (2001) *The Journal of Biochemistry* **276**, 16374
152. Okkenhaug, K., Bilancio, A., Farjot, G., Priddle, H., Sancho, S., Peskett, E., Pearce, W., Meek, S. E., Salpekar, A., Waterfield, M. D., Smith, A. J., and Vanhaesebroeck, B. (2002) *Science* **297**, 1031-1034
153. Katso, R., Okkenhaug, K., Ahmadi, K., White, S., Timms, J., and Waterfield, M. D. (2001) *Annu Rev Cell Dev Biol* **17**, 615-675
154. Suire, S., Coadwell, J., Ferguson, G. J., Davidson, K., Hawkins, P., and Stephens, L. (2005) *Curr Biol* **15**, 566-570
155. Laffargue, M., Calvez, R., Finan, P., Trifilieff, A., Barbier, M., Altruda, F., Hirsch, E., and Wymann, M. P. (2002) *Immunity* **16**, 441-451
156. Munnik, T., Irvine, R. F., and Musgrave, A. (1998) *Biochim Biophys Acta* **1389**, 222-272
157. Parker, P. J. (2004) in *Morton Medal Lecture* (Society, B., ed), pp. 893- 898, SECC, Glasgow
158. Cockcroft S. (2000) *Biology of Phosphoinositides.* Frontiers in Molecular Biology (Glover, B. D. H. a. D. M., Ed.), Oxford University Press
159. Leevers, S. J., Vanhaesebroeck, B., and Waterfield, M. D. (1999) *Curr Opin Cell Biol* **11**, 219-225
160. Franke, T. F., Kaplan, D. R., and Cantley, L. C. (1997) *Cell* **88**, 435-437
161. Sadhu, C., Masinovsky, B., Dick, K., Sowell, C. G., and Staunton, D. E. (2003) *J Immunol* **170**, 2647-2654
162. Toker, A., and Cantley, L. C. (1997) *Nature* **387**, 673-676
163. Lemmon, M. A., Ferguson, K. M., and Abrams, C. S. (2002) *FEBS Lett* **513**, 71-76
164. Lemmon, M. A., and Ferguson, K. M. (2001) *Biochem Soc Trans* **29**, 377-384
165. Vanhaesebroeck, B., and Alessi, D. R. (2000) *Biochem J* **346 Pt 3**, 561-576
166. Alessi, D. R. D., C.P. (1998) *Biochemica et Biophysica Acta* **1436**, 151-164
167. Vivanco, I., and Sawyers, C. L. (2002) *Nat Rev Cancer* **2**, 489-501
168. Shepherd, P. R. (2005) *Acta Physiol Scand* **183**, 3-12
169. Brady, M. J., and Saltiel, A. R. (1999) *J Clin Invest* **104**, 675-676
170. Clayton, E., Bardi, G., Bell, S. E., Chantry, D., Downes, C. P., Gray, A., Humphries, L. A., Rawlings, D., Reynolds, H., Vigorito, E., and Turner, M. (2002) *J Exp Med* **196**, 753-763
171. Suzuki, H., Terauchi, Y., Fujiwara, M., Aizawa, S., Yazaki, Y., Kadowaki, T., and Koyasu, S. (1999) *Science* **283**, 390-392
172. Suzuki, A., Yamaguchi, M. T., Ohteki, T., Sasaki, T., Kaisho, T., Kimura, Y., Yoshida, R., Wakeham, A., Higuchi, T., Fukumoto, M., Tsubata, T., Ohashi, P. S., Koyasu, S., Penninger, J. M., Nakano, T., and Mak, T. W. (2001) *Immunity* **14**, 523-534
173. Stokoe, D. (2005) *Expert Rev Mol Med* **7**, 1-22
174. Hirsch, E., Wymann, M. P., Patrucco, E., Tolosano, E., Bulgarelli-Leva, G., Marengo, S., Rocchi, M., and Altruda, F. (2000) *Gene* **256**, 69-81
175. Yu, J., Zhang, Y., McIlroy, J., Rordorf-Nikolic, T., Orr, G. A., and Backer, J. M. (1998) *Mol Cell Biol* **18**, 1379-1387
176. Sticht, C., Hofele, C., Flechtenmacher, C., Bosch, F. X., Freier, K., Lichter, P., and Joos, S. (2005) *Br J Cancer* **92**, 770-774

177. Ma, Y. Y., Wei, S. J., Lin, Y. C., Lung, J. C., Chang, T. C., Whang-Peng, J., Liu, J. M., Yang, D. M., Yang, W. K., and Shen, C. Y. (2000) *Oncogene* **19**, 2739-2744
178. Shayesteh, L., Lu, Y., Kuo, W. L., Baldocchi, R., Godfrey, T., Collins, C., Pinkel, D., Powell, B., Mills, G. B., and Gray, J. W. (1999) *Nat Genet* **21**, 99-102
179. Massion, P. P., Taflan, P. M., Shyr, Y., Rahman, S. M., Yildiz, P., Shakthour, B., Edgerton, M. E., Ninan, M., Andersen, J. J., and Gonzalez, A. L. (2004) *Am J Respir Crit Care Med* **170**, 1088-1094
180. Mizoguchi, M., Nutt, C. L., Mohapatra, G., and Louis, D. N. (2004) *Brain Pathol* **14**, 372-377
181. Rieusset, J., Roques, M., Bouzakri, K., Chevillotte, E., and Vidal, H. (2001) *FEBS Lett* **502**, 98-102
182. Abell, K., Bilancio, A., Clarkson, R. W., Tiffen, P. G., Altaparmakov, A. I., Burdon, T. G., Asano, T., Vanhaesebroeck, B., and Watson, C. J. (2005) *Nat Cell Biol* **7**, 392-398
183. Okamoto, T., Namikawa, K., Asano, T., Takaoka, K., and Kiyama, H. (2004) *Brain Res Mol Brain Res* **131**, 119-125
184. Spender, L. C., Whiteman, H. J., Karstegl, C. E., and Farrell, P. J. (2005) *Oncogene* **24**, 1873-1881
185. Baier, R., Bondeva, T., Klinger, R., Bondev, A., and Wetzker, R. (1999) *Cell Growth Differ* **10**, 447-456
186. Kent, W. J. (2002) *Genome Res* **12**, 656-664
187. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005) *Nucleic Acids Res* **33**, D34-38
188. Chantry, D., Vojtek, A., Kashishian, A., Holtzman, D. A., Wood, C., Gray, P. W., Cooper, J. A., and Hoekstra, M. F. (1997) *J Biol Chem* **272**, 19236-19241
189. Gruss, H. J., and Dower, S. K. (1995) *Blood* **85**, 3378-3404
190. Trama, J., Go, W. Y., and Ho, S. N. (2002) *J Immunol* **169**, 5477-5488
191. Skiba, B., Neill, B., and Piva, T. J. (2005) *Photodermatol Photoimmunol Photomed* **21**, 173-182
192. Kripke, M. L., Cox, P. A., Alas, L. G., and Yarosh, D. B. (1992) *Proc Natl Acad Sci U S A* **89**, 7516-7520
193. Boccadoro, M., Morgan, G., and Cavenagh, J. (2005) *Cancer Cell Int* **5**, 18
194. Fleming, J. A., Lightcap, E. S., Sadis, S., Thoroddsen, V., Bulawa, C. E., and Blackman, R. K. (2002) *Proc Natl Acad Sci U S A* **99**, 1461-1466
195. Geiman, T. M., and Robertson, K. D. (2002) *J Cell Biochem* **87**, 117-125
196. Jenuwein, T., and Allis, C. D. (2001) *Science* **293**, 1074-1080
197. Suzuki, H., Matsuda, S., Terauchi, Y., Fujiwara, M., Ohteki, T., Asano, T., Behrens, T. W., Kouro, T., Takatsu, K., Kadowaki, T., and Koyasu, S. (2003) *Nat Immunol* **4**, 280-286
198. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000) *Nature* **403**, 503-511
199. Tsybouleva, N., Zhang, L., Chen, S., Patel, R., Lutucuta, S., Nemoto, S., DeFreitas, G., Entman, M., Carabello, B. A., Roberts, R., and Marian, A. J. (2004) *Circulation* **109**, 1284-1291

200. Hoffmann, R., and Valencia, A. (2004) *Nat Genet* **36**, 664
201. Kippenberger, S., Loitsch, S., Guschel, M., Muller, J., Kaufmann, R., and Bernd, A. (2005) *FEBS Lett* **579**, 207-214
202. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., and et al. (1991) *Science* **252**, 1651-1656
203. Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., White, O., and et al. (1995) *Nature* **377**, 3-174
204. Clayton, E., McAdam, S., Coadwell, J., Chantry, D., and Turner, M. (2001) *Biochem Biophys Res Commun* **280**, 1328-1332
205. Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schonbach, C., Gojobori, T., Baldarelli, R., Hill, D. P., Bult, C., Hume, D. A., Quackenbush, J., Schriml, L. M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K. W., Blake, J. A., Bradt, D., Brusic, V., Chothia, C., Corbani, L. E., Cousins, S., Dalla, E., Dragani, T. A., Fletcher, C. F., Forrest, A., Frazer, K. S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I. J., Jarvis, E. D., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, R. M., King, B. L., Konagaya, A., Kurochkin, I. V., Lee, Y., Lenhard, B., Lyons, P. A., Maglott, D. R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W. J., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J. U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J. C., Reed, D. J., Reid, J., Ring, B. Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C. A., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M. S., Teasdale, R. D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, L. G., Wynshaw-Boris, A., Yanagisawa, M., Yang, I., Yang, L., Yuan, Z., Zavolan, M., Zhu, Y., Zimmer, A., Carninci, P., Hayatsu, N., Hirozane-Kishikawa, T., Konno, H., Nakamura, M., Sakazume, N., Sato, K., Shiraki, T., Waki, K., Kawai, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Miyazaki, A., Sakai, K., Sasaki, D., Shibata, K., Shinagawa, A., Yasunishi, A., Yoshino, M., Waterston, R., Lander, E. S., Rogers, J., Birney, E., and Hayashizaki, Y. (2002) *Nature* **420**, 563-573
206. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina,

N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001) *Science* **291**, 1304-1351

207. Gardiner-Garden, M., and Frommer, M. (1987) *J Mol Biol* **196**, 261-282
208. Davuluri, R. V., Grosse, I., and Zhang, M. Q. (2001) *Nat Genet* **29**, 412-417
209. Kola I, B. S., Green A, Garber R, Tymms M, Papas T and Seth A. (1993) *Proc Natl Acad Sci U S A* **90**, 7588-7592
210. Molnar, A., and Georgopoulos, K. (1994) *Mol Cell Biol* **14**, 8292-8303
211. Sytina, E. V., and Pankratova, E. V. (2003) *Mol Biol (Mosk)* **37**, 755-767
212. DeKoter, R. P., Lee, H. J., and Singh, H. (2002) *Immunity* **16**, 297-309
213. Anderson, G. M., and Freytag, S. O. (1991) *Mol Cell Biol* **11**, 1935-1943
214. Arcaro, A., Khanzada, U. K., Vanhaesebroeck, B., Tetley, T. D., Waterfield, M. D., and Seckl, M. J. (2002) *Embo J* **21**, 5097-5108
215. Felsenfeld, G. (1992) *Nature* **355**, 219-224
216. Deissler, H., Behn-Krappa, A., and Doerfler, W. (1996) *J Biol Chem* **271**, 4327-4334